Slovenská technická univerzita
Fakulta informatiky a informačných technológií
Ilkovičova 2, 842 16 Bratislava

Vyhľadávanie informácií - Konzultácia 2

Katarína Juhásová

Cvičiaci: Mgr. Martin Šeleng, PhD.
Študijný odbor: Inteligentné softvérové systémy
Ročník: 1. ročník, Ing. štúdium
Akademický rok: 2020/2021

Dáta

Authors

- informácie o autoroch
- každý riadok obsahuje id entity + určitý typ informácie, napr. *rank, name, paperCount, citationCount…*
- pre riešenie úlohy je dôležitý predovštkým parameter *name*
- pre získavanie zaujímavých štatistík môžu byť prípadne zaujímavé aj ďalšie parametre ako napr. *citationCount* alebo *paperCount*, tie by však malo byť možné dopočítať aj po správnom spárovaní autorov a článkov

```
<http://mag.graph/entity/2746065156> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://mag.graph/class/Author> .
<http://mag.graph/entity/2746065156> <http://mag.graph/property/rank> "18713"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2746065156> <http://xmlns.com/foaf/0.1/name> "Masaro Takahashi"^^<http://www.w3.org/2001/XMLSchema#string> .
<http://mag.graph/entity/2746065156> <http://mag.graph/property/paperCount> "2"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2746065156> <http://mag.graph/property/citationCount> "16"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2746065156> <http://purl.org/dc/terms/created> "2017-08-31"^^<http://www.w3.org/2001/XMLSchema#date> .

<http://mag.graph/entity/2684155564> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://mag.graph/class/Author> .
<http://mag.graph/entity/2684155564> <http://mag.graph/property/rank> "19905"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2684155564> <http://xmlns.com/foaf/0.1/name> "M. C. Hanumantharaju"^^<http://www.w3.org/2001/XMLSchema#string> .
<http://mag.graph/entity/2684155564> <http://mag.graph/property/paperCount> "1"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2684155564> <http://mag.graph/property/citationCount> "1"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2684155564> <http://purl.org/dc/terms/created> "2017-06-30"^^<http://www.w3.org/2001/XMLSchema#date> .

<http://mag.graph/entity/2034529996> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://mag.graph/class/Author> .
<http://mag.graph/entity/2034529996> <http://mag.graph/property/rank> "19760"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2034529996> <http://xmlns.com/foaf/0.1/name> "M Abbink"^^<http://www.w3.org/2001/XMLSchema#string> .
<http://mag.graph/entity/2034529996> <http://mag.graph/property/paperCount> "3"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2034529996> <http://mag.graph/property/citationCount> "8"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2034529996> <http://purl.org/dc/terms/created> "2016-06-24"^^<http://www.w3.org/2001/XMLSchema#date> .
```

Papers

- informácie o článkoch
- každý riadok obsahuje id entity + určitý typ informácie, napr. *rank, title, publicationDate, referenceCount, citationCount, estimatedCitationCount, publisher, appearsInJournal, volume…*
- pre riešenie úlohy je dôležitý predovšetkým parameter *title*
- pre získavanie zaujímavých štatistík môžu byť prípadne zaujímavé aj ďalšie parametre ako napr. *citationCount* alebo *publicationDate*

```
<http://mag.graph/entity/2811513934> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://mag.graph/class/Paper> .
<http://mag.graph/entity/2811513934> <http://mag.graph/property/rank> "22210"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2811513934> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://purl.org/spar/fabio/PatentDocument> .
<http://mag.graph/entity/2811513934> <http://purl.org/dc/terms/title> "ICE HEAT STORAGE TYPE AIR CONDITIONER"^^<http://www.w3.org/2001/XMLSchema#string> .
<http://mag.graph/entity/2811513934> <http://prismstandard.org/namespaces/1.2/basic/publicationDate> "2003-05-08"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://mag.graph/entity/2811513934> <http://mag.graph/property/citationCount> "1"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2811513934> <http://mag.graph/property/estimatedCitationCount> "1"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2811513934> <http://purl.org/dc/terms/created> "2018-07-19"^^<http://www.w3.org/2001/XMLSchema#date> .

<http://mag.graph/entity/1479662854> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://mag.graph/class/Paper> .
<http://mag.graph/entity/1479662854> <http://mag.graph/property/rank> "25119"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/1479662854> <http://purl.org/dc/terms/title> "The Hume Literature, 1994"^^<http://www.w3.org/2001/XMLSchema#string> .
<http://mag.graph/entity/1479662854> <http://prismstandard.org/namespaces/1.2/basic/publicationDate> "1995-01-01"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://mag.graph/entity/1479662854> <http://purl.org/dc/terms/publisher> "Hume Society"^^<http://www.w3.org/2001/XMLSchema#string> .
<http://mag.graph/entity/1479662854> <http://mag.graph/property/appearsInJournal> <http://mag.graph/entity/154562576> .
<http://mag.graph/entity/1479662854> <http://prismstandard.org/namespaces/basic/2.0/volume> "21"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/1479662854> <http://prismstandard.org/namespaces/basic/2.0/issueIdentifier> "2"^^<http://www.w3.org/2001/XMLSchema#string> .
<http://mag.graph/entity/1479662854> <http://prismstandard.org/namespaces/basic/2.0/startingPage> "357"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/1479662854> <http://prismstandard.org/namespaces/basic/2.0/endingPage> "366"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/1479662854> <http://mag.graph/property/referenceCount> "100"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/1479662854> <http://mag.graph/property/citationCount> "0"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/1479662854> <http://mag.graph/property/estimatedCitationCount> "0"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/1479662854> <http://purl.org/dc/terms/created> "2016-06-24"^^<http://www.w3.org/2001/XMLSchema#date> .
```

PaperAuthorAffiliations

- prepojenie článkov a autorov
- každý riadok predstavuje vzťah medzi článkom a autorom (prvá entita je článok, druhá predstavuje vzťah "creator" a posledná entita je autor) pričom platí, že jeden článok môže mať viac autorov a jeden autor môže byť tvorcom viacerých článkov

```
<http://mag.graph/entity/2514067917> <http://purl.org/dc/terms/creator> <http://mag.graph/entity/2334513227> .
<http://mag.graph/entity/2514067917> <http://purl.org/dc/terms/creator> <http://mag.graph/entity/2159674999> .
<http://mag.graph/entity/2514067917> <http://purl.org/dc/terms/creator> <http://mag.graph/entity/2133135916> .
<http://mag.graph/entity/2514067917> <http://purl.org/dc/terms/creator> <http://mag.graph/entity/2165266595> .
<http://mag.graph/entity/2527695015> <http://purl.org/dc/terms/creator> <http://mag.graph/entity/2645743283> .
<http://mag.graph/entity/2273582415> <http://purl.org/dc/terms/creator> <http://mag.graph/entity/2109773566> .
<http://mag.graph/entity/2273582415> <http://purl.org/dc/terms/creator> <http://mag.graph/entity/546406812> .
<http://mag.graph/entity/2273582415> <http://purl.org/dc/terms/creator> <http://mag.graph/entity/2163045884> .
<http://mag.graph/entity/2273582415> <http://purl.org/dc/terms/creator> <http://mag.graph/entity/1898881442> .
```

FieldsOfStudy

- informácie o študijných odboroch
- každý riadok obsahuje id entity + určitý typ informácie, napr. *rank, name, level, paperCount, citationCount*…
- pre riešenie úlohy je dôležitý predovšetkým parameter *name*
- z ďalších parametrov môže byť zaujímavé obsiahnuť v riešení *level*, ktorý hovorí o úrovni študijného odboru v stromovej hierarchii

```
<http://mag.graph/entity/74645175> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://mag.graph/class/FieldOfStudy> .
<http://mag.graph/entity/74645175> <http://mag.graph/property/rank> "11681"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/74645175> <http://xmlns.com/foaf/0.1/name> "Symbol rate"^^<http://www.w3.org/2001/XMLSchema#string> .
<http://mag.graph/entity/74645175> <http://mag.graph/property/level> "3"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/74645175> <http://mag.graph/property/paperCount> "3826"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/74645175> <http://mag.graph/property/citationCount> "54141"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/74645175> <http://purl.org/dc/terms/created> "2016-06-24"^^<http://www.w3.org/2001/XMLSchema#date> .

<http://mag.graph/entity/2775905019> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://mag.graph/class/FieldOfStudy> .
<http://mag.graph/entity/2775905019> <http://mag.graph/property/rank> "10342"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2775905019> <http://xmlns.com/foaf/0.1/name> "In silico"^^<http://www.w3.org/2001/XMLSchema#string> .
<http://mag.graph/entity/2775905019> <http://mag.graph/property/level> "2"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2775905019> <http://mag.graph/property/paperCount> "14770"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2775905019> <http://mag.graph/property/citationCount> "167950"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://mag.graph/entity/2775905019> <http://purl.org/dc/terms/created> "2018-01-05"^^<http://www.w3.org/2001/XMLSchema#date> .
```

PaperFieldOfStudy

- prepojenie článkov a študijných odborov
- každý riadok predstavuje vzťah medzi článkom študijným odborom (prvá entita je článok, druhá predstavuje vzťah "hasDiscipline" a posledná entita je študijný odbor) pričom platí, že každý článok má priradený práve jeden odbor

```
<http://mag.graph/entity/1997309520> <http://purl.org/spar/fabio/hasDiscipline> <http://mag.graph/entity/21774173> .
<http://mag.graph/entity/2385065758> <http://purl.org/spar/fabio/hasDiscipline> <http://mag.graph/entity/47768531> .
<http://mag.graph/entity/2052111847> <http://purl.org/spar/fabio/hasDiscipline> <http://mag.graph/entity/185592680> .
<http://mag.graph/entity/2212375969> <http://purl.org/spar/fabio/hasDiscipline> <http://mag.graph/entity/2776741139> .
<http://mag.graph/entity/2603556654> <http://purl.org/spar/fabio/hasDiscipline> <http://mag.graph/entity/134018914> .
<http://mag.graph/entity/2047796048> <http://purl.org/spar/fabio/hasDiscipline> <http://mag.graph/entity/63648874> .
<http://mag.graph/entity/2062994239> <http://purl.org/spar/fabio/hasDiscipline> <http://mag.graph/entity/126322002> .
<http://mag.graph/entity/2417177136> <http://purl.org/spar/fabio/hasDiscipline> <http://mag.graph/entity/86803240> .
```

FieldOfStudyChildren

- informácie o vzťahoch medzi študijnými odbormi, niektoré sú totiž hierarchicky vyššie, napr. študijný odbor “Distributive law between monads” je podriedou odboru “Distributive property”
- každý riadok predstavuje vzťah medzi 2 študijnými odbormi (prvá entita je dcérsky odbor, druhá predstavuje vzťah “hasParent” a posledná entita je rodičovský študijný odbor) pričom platí, že pod jeden rodičovský odbor môže patriť viac dcérskych odborov

```
<http://mag.graph/entity/8346251> <http://mag.graph/property/hasParent> <http://mag.graph/entity/51460> .
<http://mag.graph/entity/32952772> <http://mag.graph/property/hasParent> <http://mag.graph/entity/51460> .
<http://mag.graph/entity/2780907442> <http://mag.graph/property/hasParent> <http://mag.graph/entity/51460> .
<http://mag.graph/entity/41220328> <http://mag.graph/property/hasParent> <http://mag.graph/entity/51460> .
<http://mag.graph/entity/80672880> <http://mag.graph/property/hasParent> <http://mag.graph/entity/51460> .
<http://mag.graph/entity/8245965> <http://mag.graph/property/hasParent> <http://mag.graph/entity/51460> .
<http://mag.graph/entity/13730304> <http://mag.graph/property/hasParent> <http://mag.graph/entity/51460> .
<http://mag.graph/entity/90984967> <http://mag.graph/property/hasParent> <http://mag.graph/entity/51460> .
```

Pozn.: riadky obsahujúce type, napr. “*<http://mag.graph/entity/559973243> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://mag.graph/class/Paper>* .” nie sú až tak dôležité, keďže informácie o autoroch, článkoch, študijných odboroch (field of study) a prepojenia medzi nimi sú roztriedené do rôznych súborov podľa tohto typu a pre riešenie úlohy nie je dôležité, či ide o článok alebo knihu a pod. Zároveň aj parameter *created* nemá pre nás žiadnu výpovednú hodnotu.

Návrh riešenia

V prvom a najrozsiahlejšom kroku bude potrebné vytvoriť JSON súbor obsahujúci vybrané informácie o autoroch (id entity, meno autora a pod.) a pre každého autora jemu príslušné články

Návrh riešenia projektu je možné rozdeliť na nasledovné kroky:

1. Vyparsovanie id a mien autorov do .json súboru pomocou regex-ov
2. Prevedenie analýzy nad menami autorov, ich zoradenie podľa abecedy a zistenie, či je možné, že by sa v súbore vyskytovali rovnaké osoby pod viacerými rôznymi menami (napr. iné formáty mena)
    a. Ak sa v súbore takéto mená vyskytujú, budú zgrupené pod jedného autora, ktorému budú tým pádom prislúchať viaceré indexy entít
    b. Ak sa v súbore nenachádzajú žiadni autori pod viacerými menami, môže sa rovno prejsť k ďalšiemu kroku
3. Vyparsovanie zoznamu článkov pre každého autora, opäť pomocou regex-ov.
   K článkom budú zároveň priradené ich študijné odbory (pravdepodobne zoznam odborov predstavujúci cestu ku koreňovému odboru) a ďalšie informácie ako počet citácií, dátum vydania a pod.

4. Zistenie, či je možné, že viacej autorov vystupuje pod rovnakým menom. To by malo byť možné previesť pomocou študijných odborov jednotlivých článkov, ak sú koreňové odbory článkov pre jedného autora rôzne, je možné, že ide o rôznych autorov a záznam pre daného autora bude rozdelený na 2 alebo viac podľa rôzneho zamerania článkov
5. Nahranie vytvoreného súboru na Hadoop cluster.
6. Implementácia scriptov, využívajúcich MapReduce prístup na zistenie rôznych štatistík, napr. počet autorov publikujúcich v určitom študijnom odbore, autorov s najväčším počtom publikácií a pod.

Architektúra riešenia