Semantische connectie-machine.

Inleiding

De projectenportfolio is een verzameling van de informatie van onderzoeksgroepen en samenwerkingsverbanden gerelateerd aan de HZ.

Deze informatie wordt bewaard in een mediawiki.
Mediawiki is de onderliggende techniek van Wikipedia, een bundeling van kennis. Mediawiki is daarom van nature uit een goede manier om dergelijke kennis op te slaan, maar ook om informatie te koppelen via links en categorieen. Ook het zoeken in een mediawiki is tegenwoordig geoptimaliseerd voor zoeken in de mediawiki-kennisbank.

In de nieuwste versies van Mediawiki wordt voor het zoeken gebruik gemaakt van ElasticSearch. ElasticSearch is een tool die erg goed diverse informatiebronnen en data met verschillende formaten kan koppelen en doorzoeken. Ook kan ElasticSearch meta-data toevoegen aan intern binnen ElasticSearch opgeslagen data. Dit kan via aangemaakte indexen, maar ook door extra data aan de data-verzameling binnen ElasticSearch toe te voegen.
Dit leidt binnen ElasticSearch tot een kopie van alle data die in de mediawiki is opgeslagen.
ElasticSearch wordt up-to-date gehouden door de extensie CirrusSearch.
CirrusSearch maakt standaard een aantal indexen aan in ElasticSearch. Ook houdt CirrusSearch de data van ElasticSearch up-to-date door aangepaste pagina' s binnen mediawiki direct te updaten in de informatieverzameling van ElasticSearch. CirrusSearch biedt hooks aan als je eigen aanpassingen en data/toevoegingen wilt bewaren voor deze gewijzigde pagina's

Kenmerkend voor de projectenportfolio is dat er niet alleen data en pages worden gerelateerd, maar dat er ook gebruik gemaakt wordt van semantische data om de informatie in de projectenportfolio te structureren.
Deze extra informatie wordt door CirrusSearch standaard niet meegenomen in de zoek-indexen, bahalve als het in de platte tekst van de wiki-pages staat. Dat is jammer, want de properties "Heading (nl/en etc.)" en "Summary (en/nl/ etc.)" bevatten geconcentreerde informatie die goed bij het zoeken te gebruiken is. De redenering is dat zoek woorden die overeenkomen met woorden in de summary tot een hogere rating/ranking van de pagina in kwestie zou moeten leiden.
Ook de properties "Supercontext" en "Topcontext" hebben een semantische betekenis. Zoekresultaten gevonden in een supercontext zouden hoger in de zoekresultaten terug moeten komen dan zoekresultaten in een willekeurige pagina.
Andere properties kunnen ook betekenis hebben. De redenering is simpel: als er een verwijzing is naar een andere pagina, dan is informatie op die pagina ook sterker gerelateerd dan als er geen verwijzing is. Alleen is die koppeling afhankelijk van de property, en waarschijnlijk niet altijd zo duidelijk als bij "Supercontext" en "Topcontext"
Een andere standaard resource die bij de projectenportfolio wordt gehanteerd is de Resource Description. Die bevat soms pdf- of andere resources waarvan de inhoud dus gekoppeld is aan de pagina waar de resource description gebruikt wordt. Ook pagina's die gebruik maken van dezelfde Resource Description zijn op die manier gekoppeld.
Verder wordt er in de wiki via propeties bijgehouden welke pagina bij wel project hoort.
Omdat de onderzoeksgroepen in de wiki van elkaar onderscheiden zijn door het project-id, is het hiermee mogelijk per onderzoeksgroep/project indexen in ElasticSearch aan te maken waarmee binnen een onderzoeksgroep gezocht kan worden, of waarmee onderzoeksresultaten van een onderzoeksgroep een hogere ranking krijgen.

Er is bij de EVM in de laatste maanden een ontwikkel-omgeving gemaakt waarin de technische infrastructuur is opgebouwd om de bovenstaande elementen (zoeken, Semantische Mediawiki 2.5.8, CirrusSearch en ElasticSearch 5.6) te kunnen combineren om een zgn. semantische connectie-machine te maken. Hierbij wordt van de Elastica-extensie gebruik gemaakt om als schil te dienen tussen CirrusSearch en de API van ElasticSearch.
Binnen deze connectie-machine zal niet alleen een rangorde aangegeven worden voor de zoekresultaten, maar ook tooling gemaakt voor bijvoorbeeld een lijst met links op een pagina die gerelateerd zijn (bepaald door ElasticSeach, en daarbinnen ook bewaard)

Doel

Het doel van dit project is om uiteindelijk binnen een pagina een lijst met pagina's te kunnen laten zien die sterk gerelateerd zijn aan de huidige pagina.
Als tussenresultaten daarvoor kunnen deelprodukten gemaakt worden die:
– zoekresultaten ranken mbv. de inhoud van de "Summery", "Semantic title" – properties
– zoekresultaten ranken tov. de "Supercontext", "Topcontext" en "Resource description"-properties
– de content van de pdf-resources meenemen in de zoekresultaten
– zoekresultaten beperken tot het huidige project

Stappen

1) Installeren lokaal van testversie met werkende configuratie. Gebruikte versies: MediaWiki 1.30, SMW 2.5.8, ElasticSearch 5.6, PHP 7.1
   Hierbij wordt ervaring opgedaan (en gedeeld) die gebruikt kan worden om deze combinatie ook op de acceptatie-server en op de productie-server te installeren.
2) Op basis van de al door Anton gemaakte extensie bovenop CirrusSearch de benodigde semantische informatie door te sluizen naar shards en indexen van ElasticSearch
3) Op basis van deze data de zoekresultaten te kunnen beperken tot 1 project (toe te passen op KCKT in de projectenportfolio)
4) Meertaligheid inbouwen in de zoek-functionaliteit. Meertaligheid komt tot uitdrukking in de properties "Heading en" etc, "Summery en" etc, en de pagina's ...en/pagename etc.
5) Verrijken van de data bewaard binnen ElasticSearch via diverse python-libraries.
   Bijvoorbeeld om analyses te doen naar termen die in een bundel pagina's gebruikt kunnen worden als onderscheidende termen.
   Dit gebeurt in een aantal stappen:
   – uitfilteren onbelangrijke termen / stopwoorden
   – gebruik van synoniemen-lijsten om bewaarde teermen te standaardiseren
   – toepassen van de juiste analyse-techniek
   (Anton heeft al een voorbeeld gemaakt voor deze drie stappen)
6) Maken van een extensie / parserfunctie om dergelijke informatie binnen een pagina weer te geven.
   Mogelijke presentaties binnen een wiki-pagina:
   – tagcloud
   – een lijst met: kijk hier ook eens
   – etc.

Stap 1 t/m 3 zal binnen 4 tot 6 weken (dus uiterlijk 1 december) tot een presenteerbaar prototype moeten leiden, dat gepresenteerd kan worden aan KCKT en andere belanghebbenden.