

Significante term extractie

overzicht testen

voor	RNA-project
van	Trezorix, Aduna
datum	30-05-2007
betreft	overzicht van testen en resultaten m.b.t. significante term extractie
versie	1b

1 Inleiding

Het idee achter de "extractor" die bij onderstaande tests op gebied van significante term extractie is gebruikt, is heel simpel samen te vatten:

- sorteer alle woorden in een document op frequentie
- verwijder woorden die niet voldoen aan bepaalde minimumeisen, b.v.
- omtrent lengte
- verwijder alle woorden die in een van de stopwoordenlijsten voorkomen
- filter termen die duplicaten van elkaar lijken te zijn, b.v. als zowel "camping" en als "campings" voorkomen, zorg dan dat je er maar eentje overhoudt
- neem van de overgebleven termen de 10 meest voorkomende

In principe werkt de extractor taalafhankelijk. De stopwoordenlijsten zijn wel taalspecifiek maar ze worden in de code gemerged tot één lijst. Het idee is dat stopwoorden in de ene taal dikwijls òf niet voorkomen in een andere taal òf in die andere taal ook een stopwoord zijn. Voordeel van deze aanpak is dat we dus niet afhankelijk zijn van de aanwezigheid en kwaliteit van een language identifier. De enkele gevallen waarbij een stopwoord wel degelijk betekenisvol is in een andere taal (b.v. het Nederlandse vs. het Engelse woord "van") zijn redelijk zeldzaam en voor onze doeleinden te verwaarlozen. Het gaat ons immers vooral om het doen van zoeksuggesties en niet per se om het met een grote mate van accuratesse samenvatten van een document.

In onze implementatie maken we gebruik van tussenresultaten die gegenereerd worden door Apache Lucene. Lucene is een Open Source full-text indexer in Java die we al gebruiken voor het realiseren van de full-text search. Lucene maakt gebruik van zogenaamde Analyzers om een tekst te pre-processen (lowercasing e.d.) en te tokenizen. Door de Analyzer te wrappen in een custom analyzer die te token stream aflijstert, voorkomen we dat we voor de termextractie opnieuw de tekst hoeven te processen: we hoeven alleen maar een mapping van termen naar frequenties bij te houden.

Het filteren op duplicaten checkt op termen die max. 2 characters in lengte schelen en waarvan de lange term begint met de letters van de korte term. Mijn ervaring is dat dit vrijwel de meeste vervoegingen eruit filtert, zoals enkel- en meervoudsvormen en vervoegingen van werkwoorden. Er zijn uitzonderingen (b.v. "museum" vs. "musea") maar deze zijn voor onze doeleinden niet erg storend. Hiervoor geldt hetzelfde als voor het toepassen van de stopwoorden: het is erg prettig dat het algoritme niet afhankelijk is van enige vorm van taaldetectie. Een document dat geen of onverwachte significante termen heeft omdat de taalclassificatie ontbreekt of incorrect is, is vervelender dan een enkel woord dat er tussendoor glipt.

2 Opzet

- Er is gewerkt met een set van 100 documenten uit Natuurinformatie.nl
- De documenten zijn redelijk willekeurig gekozen. Hierbij zijn wel drie criteria gehanteerd:
 - gelijkmatige verdeling naar ouderdom van artikel;
 - representatieve selectie van onderwerpen;
 - een redelijke hoeveelheid tekstuele inhoud (zie volgende punt).
- De set van 100 artikelen bevat in totaal 40670 woorden.
 - Hiervan zijn er 19996 uniek (gemiddeld 49%, kleine artikelen hebben in verhouding meer unieke woorden).
 - 5 artikelen hebben minder dan 100 woorden (78,6% uniek),
 - 67 artikelen meer dan 100 en minder dan 500 (56,2% uniek),
 - 23 artikelen hebben meer dan 500 woorden en minder dan 1000 (47,4% uniek),
 - 5 artikelen hebben meer dan 1000 woorden (40% uniek).
- Op deze set documenten zijn 14 verschillende metadateringswijzen toegepast - zie volgende paragraaf.

3 Metadateringswijzen

De volgende metadateringswijzen worden getest:

nr	bronnen globale termen	verwerking	geëxtraheerde termen	globale termen	vrije termen
1	nvt	handmatig	maximaal 3		vrij aantal
2	nvt	handmatig	vrij aantal		vrij aantal
3	referentiestructuren	auto		maximaal 3	
4	referentiestructuren	handmatig op 3		maximaal 3	vrij aantal
5	referentiestructuren	handmatig op 3	maximaal 3		vrij aantal
6	referentiestructuren	auto		vrij aantal	
7	referentiestructuren	handmatig op 6		vrij aantal	vrij aantal
8	referentiestructuren	handmatig op 6	vrij aantal		vrij aantal
9	indexeringselectie	auto		maximaal 3	
10	indexeringselectie	handmatig op 9		maximaal 3	vrij aantal
11	indexeringselectie	handmatig op 9	maximaal 3		vrij aantal
12	indexeringselectie	auto		vrij aantal	
13	indexeringselectie	handmatig op 12		vrij aantal	vrij aantal
14	indexeringselectie	handmatig op 12	vrij aantal		vrij aantal

4 Wijze van testen

Bij het testen worden de volgende regels toegepast:

- Er worden telkens dezelfde 100 pagina's gestest.
- Bij elke metadateringswijze (behalve automatisch) wordt hiervoor de tijd gemeten.
- Allereerst worden geëxtraheerde en/of globale termen toegekend.
- Pas daarna worden persé noodzakelijk geachte vrije termen toegekend. Uitgangspunt hierbij is dat we naar de resultaten met en zonder vrije termen kunnen kijken.

5 Metadatering

5.1.1 Domeintrefwoord

Aan ieder artikel is een 'domeintrefwoord' toegekend (in uitzonderlijke gevallen meer dan 1). De hiervoor nodige domeintrefwoorden komen nauwelijks voor in de referentielijst of voorindexlijst. Ook komen deze termen nauwelijks voor in de tekst van de artikelen, dus werd het trefwoord bijna nooit geëxtraheerd. Hierdoor zijn ze meestal als 'vrije term' toegevoegd. In een enkel geval is in de globale termenlijst wel een variant of synoniem van het gewenste trefwoord te vinden. Soms is deze gekozen, maar niet altijd. (Bijv.: 'paleoantropologie' komt niet als globale term voor, wel 'paleoantropoloog'. In een deel van de sets is 'paleoantropoloog' als domeintrefwoord toegekend, in een deel is 'paleoantropologie' als vrije term toegevoegd.) In het overzicht van de resultaten wordt daarom aangegeven waar en hoeveel domeintrefwoorden zijn toegevoegd.

In de metadateringswijzen waar een maximum aantal termen (4, 5, 10, 11) geldt, tellen de domeintrefwoorden mee in het maximum als ze niet als vrije term zijn toegevoegd.

5.1.2 Voorkeursterm

Populair vs. wetenschappelijk

Als in een artikel zowel de wetenschappelijke term als de populaire term voorkomen, is de gewone term gekozen als trefwoord. In de eerste sets is dan als opmerking de bijbehorende wetenschappelijke term toegevoegd (bijv. nederlandse soortnamen als metadata toegevoegd, de wetenschappelijke naam als opmerking). In de metadateringswijzen waar geen maximum aantal termen geldt en beide varianten automatisch zijn geëxtraheerd, zijn beide vormen gehandhaafd.

Enkelvoud vs. meervoud

In de referentielijst komt van een groot aantal termen ook de meervoudsvariant voor. Geregeld zijn bij de automatische extractie zowel het enkelvoud als het meervoud van de term aangeboden. In de verwerkingswijzen waarbij een maximum geldt (4, 5, 10, 11) is de variant in enkelvoud gekozen. Bij de andere verwerkingen zijn beide varianten gehandhaafd.

Samengestelde termen

Samengestelde termen kunnen niet automatisch geëxtraheerd worden. Samengestelde termen komen ook niet voor in de voorindex. In de referentielijst komen wel samengestelde termen voor, maar ook deze kunnen niet automatisch aan de artikelen gekoppeld worden. Samengestelde termen moeten dus meestal als vrije term worden toegevoegd.

De losse woorden uit de samenstelling worden wel als zelfstandige geëxtraheerde termen aangeboden. Deze moeten in het algemeen weer worden verwijderd uit de selectie.

5.1.3 Toekenning van trefwoorden

Volgorde van termsoort

Als een term zowel in de geëxtraheerde termen als in de referentielijst voorkomt, wordt de term uit de referentielijst toegevoegd als metadatatrefwoord.

Als een term uit de referentielijst door het systeem wordt aangeboden, komt dezelfde term niet voor bij de geëxtraheerde termen. Vaak echter wordt een variant of synoniem van een term uit de referentielijst aangeboden als geëxtraheerde term (als deze variant niet voorkomt in de referentielijst). In deze gevallen zou de redacteur de voorkeursterm uit de referentielijst moeten toevoegen als metadatatrefwoord en de geëxtraheerde term moeten verwijderen. Hierbij is wel een goede kennis van de inhoud van de referentielijst nodig. Belangrijk is dat er zowel mogelijk synoniemen en varianten aan de referentielijst worden toegevoegd om de automatische toekenning zo accuraat mogelijk te laten verlopen.

In geval van twijfel over het belang van een term, worden geen termen uit de referentielijst toegevoegd. Geëxtraheerde termen zijn wel gehandhaafd in geval van twijfel (vooral in de sets waar geen maximum geldt).

Maximum

Het vastgestelde maximum van 3 trefwoorden in de sets 4, 5, 10, 11 is in het algemeen genoeg om te metadateren. In een aantal gevallen is het maximum te weinig (bijv. i003307; i003026; i003916; i004774). Een uitbreiding naar ca. 5 zou waarschijnlijk in alle gevallen voldoende zijn.

6 resultaten

Tijdsduur

Tijdsduur in minuten per metadateringswijze

1	2	3	4	5	6	7	8	9	10	11	12	13	14
105	85	nvt	90	90	nvt	80	80	nvt	60	60	nvt	90	75

Gebruik referentielijst

Het controleren of termen (of varianten, synoniemen daarvan) in de referentielijst voorkomen is bewerkelijk. Het is belangrijk dat de gebruikte referentiestructuur zo uitgebreid mogelijk is. Verder moet de redacteur een goede kennis van deze referentiestructuur hebben.

Geen maximum

Metadateringswijzen waarbij geen maximum is bepaald voor het aantal trefwoorden (2, 7, 8, 13, 14) lijken in verhouding erg bewerkelijk, alhoewel dat niet altijd uit de tijd blijkt. Niet alleen moeten dan ontbrekende trefwoorden worden toegevoegd, maar ook het teveel aan

aangeboden trefwoorden moet worden verwijderd.

Hierbij speelt de accuratesse en uitgebreidheid van de gebruikte lijst een grote rol. Vooral bij gebruik van de indexeringsselectie zonder maximum (13, 14) worden erg veel termen door het systeem aangeboden. In combinatie met geëxtraheerde termen (14) wordt dit aantal nog groter, waarbij de geëxtraheerde termen vaak (inhoudelijk) overeenkomen met termen uit indexeringsselectie. Het verwijderen van onjuiste en dubbele termen is neemt bij deze verwerkingswijze het meeste tijd in beslag.

Resultaten

setnr.	verwerking	totaal aut geëxtraheerd	redtoegekend	totaal autgeëxtraheerde globale termen	redtoegekende globale termen	rednieuw toegevoegde globale termen	globale domeintrefwoorden	globale termen met diakr. tekens	samengestelde globale termen	vrije termen	domeintrefwoorden als vrije term	samengestelde vrije termen	vrije termen met diakr. tekens	alle stoptermen	totaal red toegekende metadata trefwoorden
1	red	992	146							157				3634	303
2	red	992	207							163				3634	370
3	aut			1422	295									3634	295
4	red	nvt	nvt	1422	195	76	22	1	9	121	77	15	3	3634	316
5	red	992	26	1422	198	67	38	1	7	80	60	14	2	3634	304
6	aut			1422	1422									3634	1422
7	red	nvt	nvt	1422	1201	91	41	1	11	78	53	11	2	3634	1279
8	red	992	37	1422	1102	94	44	1	5	74	56	14		3634	1213
9	aut			1919	295									3634	295
10	red	nvt	nvt	1919	217	101	50	9		89	45	19		3634	306
11	red	992	37	1919	217	100	52	10		68	38	20		3634	322
12	aut			1919	1919									3634	1919
13	red	nvt	nvt	1919	897	99	54	7		71	35	16		3634	968
14	red	992	12	1919	882	101	25	6		57	36	16		3634	951

- **verwerking:** werkwijze, 'aut' = automatisch, 'red' = redactioneel
- **totaal aut. geëxtraheerd:** aantal automatisch geëxtraheerde termen
- **totaal aut. geëxtraheerde globale termen:** aantal automatisch geëxtraheerde termen uit de gecontroleerde termenlijst (referentielijst of voorindex).
- **red. toegekend, red. toegekend en vrije termen:** de aantallen van de redactioneel toegekende trefwoorden, **totaal toegekende metadata trefwoorden** bevat som van deze 3 kolommen.
- **red. nieuw toegevoegde globale termen:** het aantal redactioneel toegevoegde termen uit gecontroleerde lijst die niet automatisch waren geëxtraheerd.
- **globale domeintrefwoorden en domeintrefwoorden als vrije term:** het aandeel van domeintrefwoorden in het aantal globale termen of vrije termen (bij de automatisch geëxtraheerde termen kwamen geen domeintrefwoorden voor).
- **samengestelde globale termen en samengestelde vrije termen:** het aandeel van samengestelde termen in het aantal globale of vrije termen (bij de automatisch geëxtraheerde termen kwamen geen domeintrefwoorden voor).

- **globale termen met diakr. tekens** en **vrije termen met diakr. tekens**: het aandeel van termen met diakritische tekens in het aantal globale of vrije termen (bij de automatisch geëxtraheerde termen kwamen geen domeintrefwoorden voor).

Titel

Handmatige test in Excel van het betrekken van de titel in trefwoordextractie (uitgegaan van set05):

- Totaal in titels 162 termen gevonden (zonder samenstellingen; zonder stoptermen)
- Hiervan zijn er 64 identiek aan via STE toegekende trefwoorden uit de content.
- Verder zijn 36 termen onderdeel van een samenstelling.
- 47 termen zijn synoniemen of varianten van via STE toegekende trefwoorden (bijv. meervoud-enkelvoud; oude spelling-nieuwe spelling; dino-dinosaurier-dinosaurus; uitgestorven-uitsterven)
- 21 termen zijn algemeen, zouden bij werkwijze zonder maximum echter meestal wel bij toegekende trefwoorden zitten
- dus: van de 162 termen zijn 147 geschikt als trefwoord

Gecontroleerde lijsten

Twee verschillende gecontroleerde lijsten zijn gebruikt. In de sets 3 t/m 8 zijn de bestaande referentielijsten uit Natuurinformatie toegepast, in de 9 t/m 14 is gewerkt met een voorindexlijst. De voorindex is een lijst van mogelijk relevante termen redactioneel geselecteerd uit de complete index van de content.

	referentielijst set 4,5,7,8	voorindex set 10,11,13,14	totaal
aantal termen	6439	1305	7744
uniek	5965 (niet in voorindex)	831 (niet in ref.lijt)	7260 (ontdubbeld)
gedeeld			474
aut. geëxtraheerd per set	1422	1919	
totaal red. toegekende termen	2696	2213	
gem. per set	674	553	
totaal red. nieuw toegevoegd	328	401	
gem. per set nieuw toegevoegd	82	100	

7 Voorstellen voor verbetering

Technisch zouden nog verbeteringen kunnen worden aangebracht in de herkenning en extractie van globale termen:

- Het kan bij bepaalde corpi zinvol zijn om de titel mee te nemen in indexering en extractie (zie overzicht hierboven).
- Er kunnen andere ranking-methoden toegepast worden dan nu is gedaan. Bijvoorbeeld: termen in een *inleiding* telt nu in verhouding erg zwaar, nl. 5x. Hier zou ook een verhouding van bijvoorbeeld *inleiding* 3x; *body* 1x en *titel* 5x) getest kunnen worden.

- Er kan gebruik gemaakt worden van *fuzziness* en *stemming* voor de herkenning van varianten.
- Synoniemen worden nu als volwaardige termen in de lijst opgenomen en gebruikt bij extractie etc. Bij een aantal artikelen worden verschillende varianten en synoniemen van dezelfde voorkeursterm geëxtraheerd. Als het systeem alleen de bijbehorende voorkeursterm aanbiedt wordt het aantal termen verminderd en de redactie daardoor overzichtelijker. Dit is nu onmogelijk omdat niet duidelijk is wat een synoniem is en wat een voorkeursterm
- Als een term al wordt aangeboden in de gecontroleerde lijst, dan zou die niet meer aangeboden moeten worden als geëxtraheerde term.
- Termen met leestekens (bijv. zoëchoor) worden door het systeem niet herkend.
- Samengestelde termen worden nooit herkend.
- Tijdens het redactieproces zouden de globale lijst en stoplijst bewerkt moeten kunnen worden. Met die gewijzigde lijst zou dan ook de extractie opnieuw moeten plaatsvinden.
- Uitbreiding van de referentiestructuur (bijv. meer spellingsvarianten: 'paddenstoel', meer synoniemen: 'dino', 'plaatteconiek') leidt ook tot een betere extractie.
- Als de meervoudsvorm in de tekst voorkomt en wordt aangeboden als mogelijke globale term, wordt toch indien aanwezig de enkelvoud uit de globale lijst geselecteerd (zie bijv. set5 - i000223, 'flagellen' aangeboden, 'flagel' geselecteerd). Deze worden dan als nieuwe term geteld. (NB niet altijd helemaal consequent, bij de sets zonder maximum zijn alle varianten van een term die worden aangeboden gehandhaafd).
- Er worden algemene termen uit de tekst geëxtraheerd (als extracted of globale term), waarvan een meer precieze vorm in de globale termenlijst beschikbaar is. Als uit de context blijkt dat deze specifieke vorm bedoeld wordt, wordt de precieze vorm gekozen (bijv. set 5 - i002057: 'collectie' in tekst en aangeboden, uit de lijst 'museumcollectie' geselecteerd). Deze gevallen worden als een nieuwe term geteld (in tabel kolom '*red. nieuw toegevoegde globale termen*').
- Eventuele homoniemen zijn niet te herkennen en daar kon dus ook geen rekening mee gehouden worden.