Practicum 1 Opzet DAR

Gerben Aalvanger
3987051
Universiteit Utrecht

Sam van der Wal
3962652
Universiteit Utrecht

Docent:
Hans Phillipi
6 mei 2014

1 Metadatabase configuratie

1.1 IDF$_k$

Bereken voor elk attribuut(A$_k$ berekenen we de IDF$_k$ voor elke value v in A$_k$, h is de bandbreedte parameter en n is het totaal aantal attributen. Voor numerieke attributen berekenen we dit met de formule: (1)

$$IDF_k(v) = log(\frac{n}{\sum_i^n e^{-\frac{1}{2}(\frac{v_i - v}{h})^2}}) \tag{1}$$

Voor gewone attributen berekenen we de IDF met de formule: (2) waarbij F$_k$(v) het aantal tuples in R is, met A$_k$ = v

$$IDF_k(v) = log(\frac{n}{F_k(v)}) \tag{2}$$

Voor elke tupel voegen we voor elke attribuut A$_k$ een nieuw attribuut IDF$_k$ aan de metadatabase toe met de waarde van de IDF van het attribuut.

1.2 QF

QF kunnen we bereken met behulp van(3)

$$\frac{RQF(v)}{RQFMax} \tag{3}$$

Hierin is RQF(v) de frequentie van waarde v in alle queries en RQFMAX de frequentie van de meest genoemde waarde in de queries. De QF slaan we net als de IDF op door een attribuut toe te voegen in de metadatabase. Dit attribuut bevat de waarde van de QF van de value van het oorspronkelijke attribuut. Deze QF wordt berekent aan de hand van een tabel die we aan de hand van de metadata aanmaken. Deze tabel kan later ook gebruikt worden voor de Jaccard coefficient. In de tabel zijn n kolommen van de oorspronkelijke attributen. Er zijn Q rijen die elk een querie representeren van de workload. Er wordt een extra attribuut toegevoegd die aangeeft hoevaak een querie is uitgevoerd.
Voor de numerieke waarden moeten we iets bedenken.

1.3 Jaccard

De tabel die opgebouwd is kunnen wij ook gebruiken om de jaccardcoefficient te berekenen, er hoeft niet verdere preprocessing plaats te vinden.

2 Query verwerken

2.1 IDF

Voor een query ranken wij door allereerst voor elk gegeven attribuut de similaritycoefficient te berekenen op basis van de IDF. Voor nummerieke attributen gebruiken wij de formule: (4)

$$S_{num}(v,q) = e^{-\frac{1}{2}(\frac{v-q}{h})^2} IDF(q) \tag{4}$$

Voor een enkel categorisch attribuut gebruiken wij de formule: (10)

$$S_{cat}(v,q) = \begin{cases} \text{Als q = v} & IDF(v) \\ \text{Anders} & 0 \end{cases} \tag{5}$$

Voor een enkel categorisch attribuut met meerdere mogelijke attribuutwaarden gebruiken wij de formule: (6)

$$S_{cat+}(v,q) = \begin{cases} \text{Als q in v} & IDF(v) \\ \text{Anders} & 0 \end{cases} \tag{6}$$

Voor een volledige tupel is nu de similarity: (7)

(7)

2.2 QF

Voor de similaritycoefficient berekenen kunnen we met QF gebruik maken van de formule: (8)

$$S(v,q) = \begin{cases} \text{Als q = v} & QF(v) \\ \text{Anders} & 0 \end{cases} \tag{8}$$

2.3 JACCARD

Jaccard is te gebruiken in combinatie met IDF of/en QF, we veranderen de similaritycoefficient naar:

$$S(v,q) = J(W(t), W(q))QF(q) \tag{9}$$

2.4 combinaties

De combinatie van QF, IDF en JACCARD levert de volgende formule op:

$$S(v,q) = J(W(t), W(q))QF(q)IDF(q) \tag{10}$$

hiervoor geld dat de