Promovendus Henk Ernst Blok ontwikkelt methode voor 'set-based' IR-query

Pleidooi voor lichtgewicht database-kernel

Robbert Hoeffnagel

De huidige data-aanwas, niet alleen in omvang, maar vooral in in diversiteit, lijkt amper bij te houden. Ook een met multimedia-functies 'opgetuigd' dbms schiet tekort, meent Henk Ernst Blok van de Universiteit Twente. Hij ging terug naar de basis en wijdde er zijn promotie-onderzoek aan. Blok pleit voor een lichtgewicht database-kernel, die naar behoefte is uit te breiden.

Kijk naar een digitale bibliotheek. Of neem een systeem voor document management. In beide gevallen gaat het om een digitalisering van voorheen in analoge vorm beschikbare informatie. Dit proces kan in meerdere stappen worden volbracht. In eerste instantie zal weliswaar een digitalisering van de basisdocumenten plaatsvinden, maar gebeurt het zoeken 'in' de content aan de hand van toegekende steekwoorden als onderwerp of auteur. In een vervolgstap worden mogelijkheden gecreëerd om daadwerkelijk in de digitale content zelf te zoeken.

In dit soort gevallen zou het erg handig zijn als de gestructureerde zoekmechanismen en bijvoorbeeld de query-optimalisatie van de database gecombineerd zouden kunnen worden met de technieken die ontwikkeld zijn in de wereld van de *information retrieval*. IR geeft aan documenten een 'ranking' mee in relatie tot de door de gebruiker ingegeven zoeksleutel. Die integratie is mogelijk, meent Henk Ernst Blok, maar is niet zonder slag of stoot te realiseren.

'ATOMIC OPERATORS'

Er bestaan verschillende manieren om te komen tot een gecombineerde manier van bevragen van gestructureerde gegevens en content data. Eén optie is het koppelen van een dmbs en een systeem voor information retrieval (IR) op het niveau van de applicatie. Hierbij draagt de toepassing zorg voor het samenbrengen van de resultaten die door beide zoekmethoden worden opgeleverd.

Een andere mogelijkheid is het implementeren van de functionaliteit van het IR-systeem in de door het dbms gebruikte querytaal. Ook is het denkbaar dat het extensiemechanisme van een modern dbms wordt gebruikt om IR-functionaliteit aan dit dmbs toe te voegen. Maar Blok koos ervoor de mogelijkheden van nog een andere aanpak te onderzoeken. Hij bekeek het gebruik van zogeheten atomic operators op de drie lagen van de dmbs-architectuur.

BESTE AANPAK

Blok noemt in zijn onderzoek diverse redenen waarom hij denkt dat dit de beste aanpak is. Die hebben alle te maken met de fundamentele verschillen tussen dbms- en IR-technologie.

Een eerste reden is gelegen in het feit dat makers van data-

Optimalisatie-aspecten 'information retrieval'

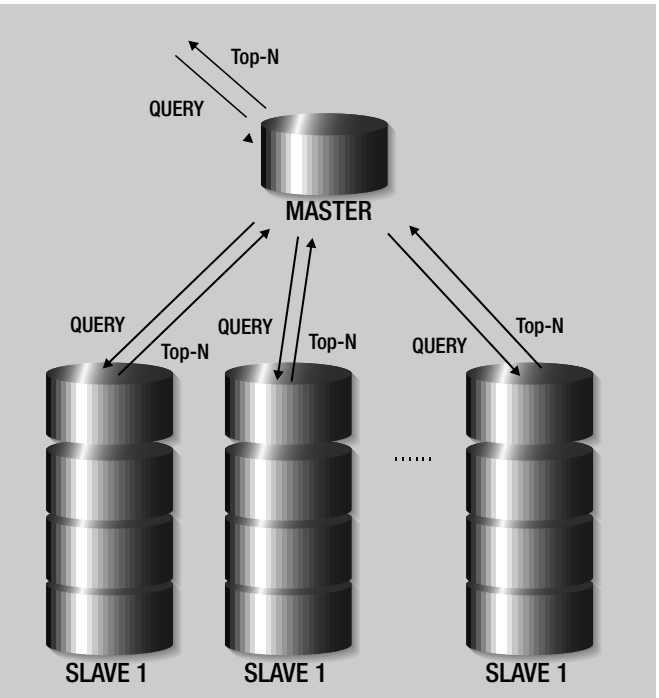
Het is een bekend gegeven: de hoeveelheid data die wordt opgeslagen groeit niet alleen in een moordend tempo, ook de diversiteit van de vastgelegde gegevens neemt steeds verder toe. Leveranciers van dmbs'en proberen hierop in te spelen door multimediale faciliteiten aan hun producten toe te voegen. Maar is dat wel voldoende, vraagt promovendus Henk Ernst Blok zich af in het onderzoek "Database optimization aspects for information retrieval" af. Een dergelijk product schiet veelal hetzij tekort in de geboden functionaliteit of kent een te gering prestatieniveau, of beide is het geval.
Blok beschouwde het proces van het verzamelen en opslaan geheel opnieuw en keek naar de combinatie van gestructureerde gegevens en contentgegevens, met name tekst. Integratie met systemen voor information retrieval is mogelijk, maar niet zonder meer. Bloks fundamentele blik mondt uit in een pleidooi voor een lichtgewicht database-kernel, waaraan afhankelijk van de behoefte die de gebruiksomgeving stelt, extra functionaliteit kan worden toegevoegd.
Dr. H.E. Blok is verbonden aan de Database Groep van de Universiteit van Twente, waar hij werkte aan het inmiddels afgeronde project voor Advanced Multimedia Indexing and Searching (AMIS).

basetechnologie er vanuit gaan dat gewerkt kan worden in een volledig gesloten omgeving. Een query wordt daarbij geacht exact te zijn. Query's in een IR-omgeving daarentegen zijn in wezen niet veel meer dan een representatie van de informatiebehoefte van de gebruiker in de vorm van steekwoorden. In tegenstelling tot de situatie bij een dbms zijn de juistheid en de kwaliteit van het resultaat van de zoekactie afhankelijk van de beoordeling van diezelfde gebruiker. Dat wordt in de literatuur ook wel *imprecise* of *inexact query processing* genoemd. Bij een combinatie van IR en dmbs zal men de laatste moeten vertellen hoe hij met dit *imprecise behaviour* moet omgaan. Veelal wordt hiervoor de Booleaanse methode gebruikt; maar die wordt inmiddels niet langer als een geschikte methode gezien, want te grof.

DATA-ONAFHANKELIJK

Daarnaast zijn IR-algoritmen veelal datagerelateerd. Dat wil zeggen dat zij bedoeld zijn voor het werken met gegevens die op een specifieke manier zijn vastgelegd. Dat is uiteraard in tegenspraak met de aanpak die in de reguliere dbms'en wordt gevolgd en waarbij juist data-onafhankelijkheid van de gebruikte algoritmen voorop staat.

Nu is het ook mogelijk het gehele IR-algoritme als een soort 'black box fysieke operator' te implementeren. Dat kan, meent Blok, maar het levert in zijn ogen wel de nodige performance-problemen op als het gaat om de optimalisatie van de query. Het IR-algoritme zal op data-onafhankelijke wijze geïmplementeerd moeten worden. Zo'n implementatie is al eerder gedaan, maar dan



FIGUUR 1: SCHEMATISCHE WEERGAVE VAN EEN PARALLELLE ARCHITECTUUR DIE DE TECHNIEKEN VAN BLOK GEBRUIKT. KUN JE IR-TOP-N-QUERY'S OPTIMALISEREN OP EEN DATABASE-MANIER? HOOGSTWAARSCHIJNLIJK WEL.

FOTO: ROBERT MUURLINK

DE OP 12 APRIL IN ENSCHEDE GEPROMOVEERDE HENK ERNST BLOK: INTEGRATIE IS MOGELIJK, MAAR NIET ZONDER SLAG OF STOOT.

vanuit het oogpunt van de verwerking van query's en niet met het oog op het verbeteren of uitbreiden van de faciliteiten van een dbms voor query-optimalisatie.

Een derde reden is dat query-optimalisatie op basis van enerzijds een herordening van operatoren en anderzijds een kostenmodel een standaard faciliteit van een rdbms is, terwijl een dergelijk mechanisme bij information retrieval in feite niet bestaat. Optimalisatie speelt bij IR natuurlijk wel een rol, maar dan vooral in de zin van betere en snellere algoritmen of algoritmen om te komen tot een betere verspreiding van data over de schijf.

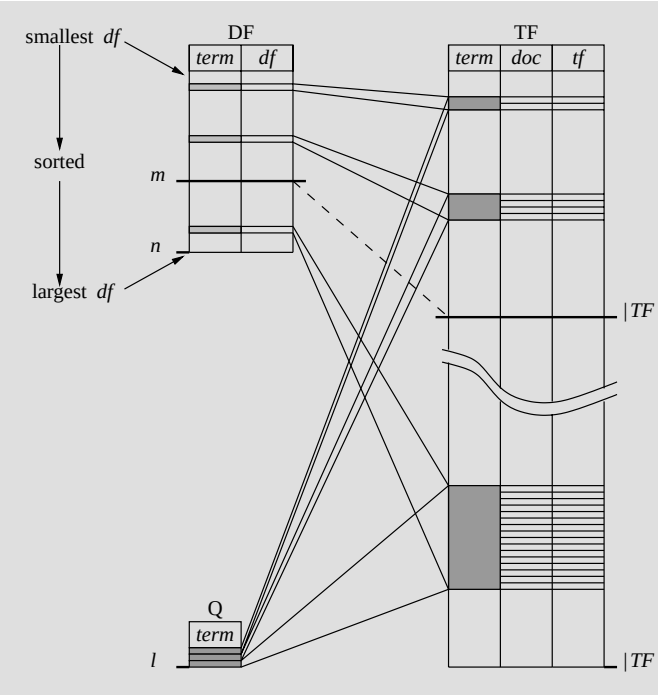
Een laatste probleem dat Blok signaleert heeft te maken met het

> *Of de retrieval-functionaliteit schiet tekort of de query-optimizer ontbreekt*

feit dat bij IR- query-verwerking sprake is van een opvolgende reeks van bevragingen en presentaties van resultaten. Bij een dbms is echter juist sprake van ad hoc query's. Zelfs als gebruik wordt gemaakt van faciliteiten als multi-query-optimalisatietechnieken voor de volgorde van query's of het parallel uitvoeren ervan, bestaat geen duidelijke relatie tussen de individuele query's. Bij information retrieval is dat nu juist wel het geval.

EIGEN AANPAK

Blok werkt in zijn onderzoek een aanpak uit die luistert naar de naam "probabilistic IR with optimizer". Daarmee probeert hij een oplossing te vinden voor een belangrijk probleem dat samenhangt met de integratie van IR- en dbms-technologie: of de retrieval-functionaliteit schiet tekort of de query-optimizer ontbreekt.



FIGUUR 2: VIRTUALISATIE VAN HET IN HET AMIS-PROJECT GEBRUIKTE DATABASESCHEMA (IN ABSTRACTO).

Gecombineerd dbms/IR-gebruik vraagt echter om een faciliteit waarmee de gebruiker de snelheid van zoeken en de kwaliteit van het zoekresultaat tegen elkaar kan afwegen.

Daarmee is ook direct het hoofddoel van het promotie-onderzoek van Blok duidelijk: ontwikkel een efficiënte en 'set-based' methodiek voor IR-query-verwerking. Kan een 'top-N IR-query' -denk aan de bekende top tien-lijstjes die zoekmachines aanbieden en die suggereren dat het hier om de meest relevante zoekresultaten gaat- worden geoptimaliseerd op een manier zoals we die kennen uit de databasetechnologie?

Formeel leidt dit tot drie onderzoeksvragen:

- Kunnen we selectiviteit schatten als een functie van het gebruikte deel van de gegevens?
- Kunnen we *quality behaviour* schatten als functie van het gebruikte deel van de data, zodat de gebruiker de genoemde afweging tussen snelheid en kwaliteit kan maken?
- Is horizontale fragmentatie van de beschikbare gegevens te gebruiken om deze afweging van snelheid versus kwaliteit te maken, zodat sprake is van de genoemde 'set-based IR top-N' query-optimalisatie?

VRAGEN BEANTWOORDEN

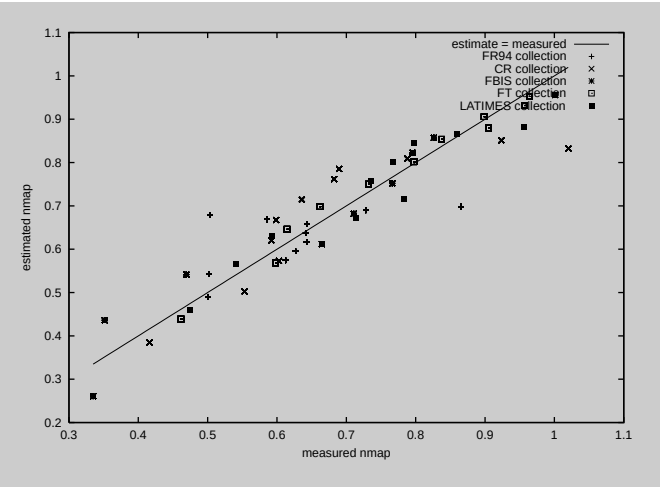
Blok heeft zich vierenhalf jaar bezig gehouden met onderzoek dat moest leiden tot het beantwoorden van deze vragen. Een deel van deze tijd werd gestoken in de bij nader inzien niet eenvoudige porting van IR-optimalisatietechnieken naar een database-omgeving.

Om de derde vraag te kunnen beantwoorden zijn vervolgens vier series van experimenten gedaan, waarbij iedere keer de executietijd is vastgesteld en twee kwaliteitsmaatstaven zijn gehanteerd (*recall* en gemiddelde nauwkeurigheid). Hierbij werd duidelijk dat het gebruiken van een kleiner eerste fragment of het voortijdig afkappen van de query-verwerking niet alleen de uitvoeringskosten verlaagt, maar tevens de kwaliteit omlaag brengt. Het

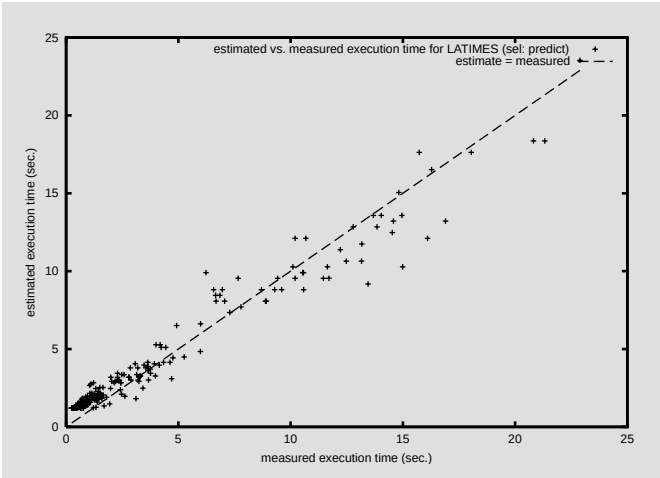
Bij gecombineerd dbms/IR-gebruik wil de gebruiker de snelheid van zoeken en de kwaliteit van het zoekresultaat tegen elkaar kunnen afwegen

gebruik van horizontale fragmentatie is wel degelijk toe te passen om voor de gebruiker een 'trade-off' tussen snelheid en kwaliteit mogelijk te maken.

Dan de vraag of de selectiviteit kan worden geschat als functie van de gebruikte fragmenten. Blok geeft aan dat bij zijn onderzoek alleen gekeken is naar het gebruik van twee fragmenten. Bij dat aantal fragmenten blijkt deze vraag positief te beantwoorden.



FIGUUR 3: GESCHATTE VERSUS GEMETEN RESULTAATKWALITEIT VOLGENS HET EERSTE KWALITEITSMODEL.



FIGUUR 4: GESCHATTE VERSUS GEMETEN EXECUTIETIJD BIJ GEBRUIKMAKING VAN BLOKS SELECTIVITEITSMODEL.

Hoewel de vraag niet in algemene zin -dus bij gebruik van grotere aantallen fragmenten- kan worden beantwoord, ziet Blok wel goede mogelijkheden dat ook in dat soort gevallen een positief antwoord mogelijk is. Maar of deze verwachting klopt, zal moeten blijken uit aanvullend onderzoek.

Resteert ten slotte de tweede vraag: kunnen we de gevolgen voor de kwaliteit van het antwoord schatten als functie van de relatieve grootte van de gebruikte fragmenten? Hiertoe heeft Blok twee modellen ontwikkeld om tot een voorspelling van het kwaliteitsgedrag te kunnen komen. Het tweede model was hierbij een generalisatie van de eerste (zie figuur 3). Eén van de modellen bleek, zo concludeert hij, *quite well* in staat om een voorspelling van het kwaliteitsgedrag te doen. Hoewel het hier in de visie van Blok slechts gaat om een eerste poging tot voorspellen, en er nog volop verbeteringen mogelijk zijn, is ook deze vraag nu al positief te beantwoorden.

ARCHITECTUUR AANPASSEN

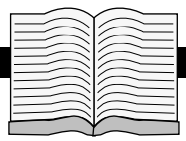
Wat zegt dit alles nu over de oorspronkelijke vraag: kunnen IR-top-N-query's worden geoptimaliseerd op een database-manier? Blok concludeert dat op basis van de antwoorden op de drie onderzoeksvragen ook deze vraag hoogstwaarschijnlijk met 'ja' beantwoord kan worden. Wel adviseert hij nader onderzoek, waarvoor hij in zijn onderzoek bovendien voorstellen doet.

Eén van die voorstellen heeft betrekking op een aantal architecturele aanpassingen die doorgevoerd zouden moeten worden om de resultaten van het promotie-onderzoek in de praktijk te kunnen gebruiken in dmbs'en met een geïntegreerde IR-functionaliteit. Steeds meer commerciële dbms'en worden door hun leveranciers uitgebreid met mogelijkheden om met multimediale content om te gaan. Steevast treedt daarbij echter het probleem op dat hetzij de functionaliteit of het prestatieniveau tekort schiet.

Aanbrengen van patches en uitbreidingen werkt niet

De oorzaak zoekt Blok hierin: men handhaaft de bestaande architectuur en probeert extra functionaliteit daaraan toe te voegen in de vorm van black boxes. Aanbrengen van patches en uitbreidingen werkt echter niet, meent hij. Blok is veel eerder voorstander van een aanpak waarbij de database-kernel zelf kleiner wordt gemaakt. Hij spreekt van een *light-weight kernel* waaraan, afhankelijk van de eisen die door de gebruiksomgeving worden gesteld, extra functionaliteit kan worden toegevoegd. ●

Robbert Hoeffnagel is freelance journalist.



Congressen, beurzen e.d.

10/10: Cognos Enterprise 2002
Relatiedag van deze BI-leverancier. Amsterdam, the Factory. Org./inf.: www.cognos/enterprise2002.

10-11/10: OGh Jubileumcongres
Oracle-gebruikersgroep, derde lustrum. Maastricht, Crowne Plaza. Org./inf.: www.oracle-usergroup.nl

21-24/10: IDUG 2002
Europese conferentie voor DB2-gebruikers. Lissabon. Org./inf.: www.idug.org

28-29/10: Data Mining Summit 2002
Congres van leverancier SPSS. Parijs, Le Meridien Etoile. Org./inf.: www.dataminingsummit.com

Cursussen, seminars e.d.

11, 18 en 25/9, 2 en 9/10: Make the smart move with XML
Wekelijkse workshop. Org.: Software AG, info: www.thexmlcompany.nl.

16-17/9: IDC Europees IT Forum
Keynotes, marktanalyses, jaaroverzicht, prognoses 2003. Monaco, Grimaldi Forum. Info: http://emea.idc.com/itforum02

19/9: Vergelijkend onderzoek CRM-software
Informatiemiddag n.a.v. rapport. Org./inf.: vandijkconsulting.com, www.ipl.nl.

7,8 en 14/10: Dimensionaal modelleren
Cursus met Harm van der Lek. Amsterdam ZO, Planetarium Gaasperplas. Org./inf.: VanderLek Advies BV, www.vdlek.nl, (035) 6216928.

15-17/10: Informatie-analyse en logisch database-ontwerp
Seminar met Rick van der Lans. Gent (B), Holiday Inn Gent Expo, 9.00-17.00 uur. Kosten: € 1350. Org./inf.: I.T. Works, www.itworks.be/logdbontwerp.html, (00) 32 9 2415613.

Alle vermelde bedragen zijn excl. BTW.

23/10: Enterprise portals
Seminar met Peter Hinssen e.a. Diegem (B), Hotel Sofitel Brussels Airport, 14.00-21.00 uur. Kosten: € 545. Org./inf.: I.T. Works, www.itworks.be/, (00) 32 9 2415613.

6/11: Ervaringen in datawarehousing
Seminar. Diegem (B), Hotel Sofitel Brussels Airport, 14.00-21.00 uur. Kosten: € 545. Org./inf.: I.T. Works, www.itworks.be/, (00) 32 9 2415613.

7/11: Enterprise applicatie-integratie
Seminar met Peter Hinssen e.a. Diegem (B), Hotel Sofitel Brussels Airport, 14.00-21.00 uur. Kosten: € 545. Org./inf.: I.T. Works, www.itworks.be/, (00) 32 9 2415613.

11-12/11: Ontwerpen van de nieuwe generatie datawarehouses
Masterclass met Rick van der Lans. Leiden, Holiday Inn, 9.30-17.00 uur. Kosten: € 1250 (€ 1175 voor DB/M-abonnees). Org.: Array Publications, info: www.array.nl, (036) 5409111.