(12) **PATENT**

(19) NO (11) **327323** (13) **B1**

NORGE

(51) Int Cl.
G06F 17/30 (2006.01)
G06F 17/00 (2006.01)

Patentstyret

(21)	Søknadsnr	20070718	(86)	Int.inng.dag og søknadsnr
(22)	Inng.dag	2007.02.07	(85)	Videreføringsdag
(24)	Løpedag	2007.02.07	(30)	Prioritet
(41)	Alm.tilgj	2008.08.08		
(45)	Meddelt	2009.06.08		

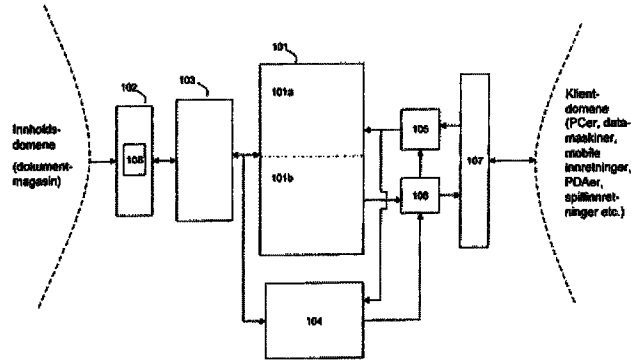
(73) Innehaver Fast Search & Transfer ASA, Postboks 1677 Vika, 0120 OSLO
(72) Oppfinner Petter Moe, Nordengkollen 71, 1396 BILLINGSTAD
(74) Fullmektig

(54) Benevnelse **Fremgangsmåte til å danne grensesnitt mellom applikasjoner i et system for søking og gjenfinning av informasjon**

(56) Anførte publikasjoner US 20040044659
US 20060074881

(57) Sammendrag

I en fremgangsmåte for å danne grensesnitt mellom applikasjoner for søking, analyse og rapportering i et system for søking og gjenfinning av informasjon i en database som rommer komplekst strukturerte dataposter eller innhold, blir en skjemaoppdagelse utført på basis av en søkeapplikasjon, skjemaveier forbundet med et søkeresultat finnes og sammendragsinformasjon for de funne skjemaveier beregnes.



Oppfinnelsen angår en fremgangsmåte til å danne grensesnitt mellom applikasjoner for søking, analyse og rapportering i et system for søking og gjenfinning av informasjon med strukturerte dokument- eller innholdsmagasiner som inneholder komplekse, strukturerte dokumenter eller innhold, hvor magasinet er søkbart og omfatter skjemaveier for dokument- og innholdsattributter og hvor fremgangsmåten omfatter trinn for å benytte et søkespørsmål for én eller flere attributtverdier på en indeks av attributtverdier, gjenfinne en resultatmengde av dokumenter eller innhold som tilsvarer nevnte én eller flere attributtverdier.

Den foreliggende oppfinnelse angår mer bestemt søkeapplikasjoner i bedrifts- eller foretakssøkesystemer, og for å belyse dette skal nå en søkemotor som kjent i teknikken og benyttet i bedriftssøkesystemer kort drøftes med henvisning til fig. 1. En søkemotor 100 som benyttet med den foreliggende oppfinnelse, omfatter som kjent i teknikken forskjellige undersystemer 101-107. Søkemotoren kan aksessere dokument- eller innholdsmagasiner plassert i et innholdsdomene eller -rom hvorfra dokumenter eller innhold enten kan aktivt skyves inn i søkemotoren eller via en datakobler trekkes inn i søkemotoren. Typiske magasiner omfatter databaser, kilder gjort tilgjengelig via ETL(Extract-Transform-Load)-verktøy så som Informatica, ethvert XML-formatert magasin, filer fra filtjenere, filer fra vevtjenere, dokumenthåndteringssystemer, innholdshåndteringssystemer, epost-systemer, kommunikasjonssystemer, samarbeidssystemer og rike media så som audio, bilder og video. De gjenfunne dokumenter leveres til søkemotoren 100 via innholds-API (Application Programming Interface) 102. Deretter blir dokumenter analysert i et innholdsanalysetrinn 103, også betegnet som et undersystem for innholdsforbehandling, for å forberede innholdet for forbedrede søke- og oppdagelsesoperasjoner. Typisk er utgangen fra dette trinn en XML-representasjon av inngangsdokumentet. Utgangen fra innholdsanalysen benyttes til å mate kjernesøkemotoren 101. Kjernesøkemotoren 101 kan typisk være anbrakt over en tjenerfarm på en desentralisert måte for å gjøre det mulig å prosessere store dokumentmengder og høye spørsmålsbelastninger. Kjernesøkemotoren 101 kan motta brukeranmodninger og frembringe lister over tilsvarende dokumenter. Dokumentrekkefølgen blir vanligvis bestemt i henhold til en relevansmodell som måler den sannsynlige betydning av et gitt dokument relativt til spørsmålet. I tillegg kan kjernesøkemotoren 103 frembringe ytterligere metadata for resultatmengden så som sammendragsinformasjon for

dokumentattributter. Kjernesøkemotoren 101 omfatter i seg selv ytterligere undersystemer, nemlig et indekseringsundersystem 101a for nedsamling ("crawling") og indeksering av dokumenter eller innhold og et søkeundersystem for å utføre egentlige søk og gjenfinning. Alternativt kan utgangen fra innholdsanalysetrinnet 103 mates inn i en valgfri varslingsmotor 104. Varslingsmotoren 104 har lagret en mengde av spørsmål og kan bestemme hvilke spørsmål som ville ha akseptert den gitte dokumentinnmating. En søkemotor kan aksesseres fra mange forskjellig klienter eller applikasjoner som typisk kan være mobile og datamaskinbaserte klientapplikasjoner. Andre klienter innbefatter PDAer og spillinnretninger. Disse klientene som befinner seg i et klientrom eller -domene vil levere anmodninger til en søkespørsmåls- eller klient-API (Application Programming Interface) 107. Søkemotoren vil typisk ha et ytterligere undersystem i form av et søkespørsmålsanalysetrinn 105 for å analysere og forfine søkespørsmålet med tanke på å konstruere et avledet søkespørsmål som kan ekstrahere mer meningsfylt informasjon. Endelig blir utgangsdata fra kjernesøkemotoren 103 typisk ytterligere analysert i et annet undersystem, nemlig et resultatanalysetrinn 106 for å frembringe informasjon eller visualiseringer som benyttes av klientene. – Begge trinnene 105 og 106 er forbundet mellom kjernesøkemotoren 101 og klient-API 107 og i tilfelle varslingsmotoren foreligger, er den forbundet i parallell til kjernesøkemotoren 101 og mellom innholdsanalysetrinnet 103 og trinnene 105;106 for henholdsvis søkespørsmåls- og resultatanalyse.

For den foreliggende oppfinnelses formål vil begrepene dokument benyttes synonymt med post, som vil brukes til å betegne objektene som utgjør en database, slik at man unngår bibetydningen av et dokument som utelukkende en tekststørrelse. Videre skal i et bedriftsmiljø en viss omfattende dokumentmengde heretter primært betraktes som en database og denne databasen er ikke bare strukturert, men også dokumentene deri vil selv være strukturert eller til og med ha en kompleks struktur. Dette står i sterk kontrast til dokumentmagasinet som påtreffes i åpne systemer så som World Wide Web, i hvilke informasjon er tilgjengelig fra et enormt antall meget forskjellige kilder og hvor informasjonsleverandørene utgjør en svært heterogen masse. Dessuten er mye av denne informasjon utstrukturert og foreligger i form av enten tekstdokumenter eller forskjellige rike media så som audio og video, som velkjent for brukerne av World Wide Web.

Fra US publisert patentsøknad nr. 2004/0044659 A1 (Judd & al.) er det kjent en fremgangmåte for søking og gjenfinning av dokumenter som tillater søking av fritekst innenfor avsnitt av skjemauavhengige dokumenter. Dokumentene kan være strukturerte, semistrukturerte og ustrukturerte og inneholde tekst som er organisert i en rekke avsnitt. Dokumentmagasinet er skjemauavhengig slik at søkesystemet ikke behøver forhåndsdefinerte søkefelt for avsnittene. I et søk mottar søkesystemet et søkespørsmål som spesifiserer minst ett avsnitt og minst et søkespørsmål konstruert som fri tekst for teksten innenfor denne seksjonen. Søkesystemet i dentifiserer avsnitt i dokumentmagasinet som angitt i søkespørsmålet og vurderer fritekstkonstruksjonen av søkespørsmålet for teksten innenfor avsnittet for å bestemme hvorhvidt begingelsen for fritekstsøking er oppfylt.

Videre angår US publisert patentsøknad nr. 2006/0074881 A1 (Vembu & al.) en framgangsmåte for å foreta stikkordbaserte søk i både strukturerte og ustrukturerte databaser og over flere databaser hos forskjellige informasjonsleverandører. Det dannes en indeksdatabase som inneholder ordforekomster og informasjon om relasjoner mellom tabeller. Dette skjer ved hjelp av en såkalt forplantende n-nivås indekseringsmetode og gjør at det kan lagres informasjon om forekomsten av ord eller relasjoner mellom nøklene til forskjellige tabeller, primær nøkkelinformasjon for alle tabeller og endelig informasjon om tabellenes rang. Indeksdatabasen dannes spesifikt ved at det benyttes et eksisterende skjema som er kjent fra strukturerte databaser og dermed kan utstrukturerte databaser som ikke har et slikt forhåndsbestemt skjema, indekseres.

Ingen av disse ovennevnte fremgangsmåter i henhold til kjent teknikk har særlig egnet til å konfigurere indekser for strukturerte databaser eller hvor databasen i seg selv omfatter dokumenter som i seg selv har en meget kompleks intern struktur.

Innenfor konteksten av en bedrift eller et foretak kan informasjon som genereres eller eies av bedriften være spredt i én eller flere databaser som typisk er fordelt over en rekke lagringsinnretninger og administrert av tjenerne til bedriften som dessuten vil støtte og betjene hvilke som helst klientgenererte applikasjoner i bedriften. Databasene er vanligvis strukturert, og i tillegg har de lagrede dokumenter i seg selv vanligvis en meget kompleks intern struktur. Et typisk eksempel ville være dokumenter som

omfatter tabeller eller lister med en blanding av numerisk informasjon og tekstinformasjon og med et stort antall attributter som er tilordnet til like store og til og med enda større strukturelle elementer av dokumentene. Tabellene og attributtene kan anses å utgjøre en informasjonsmengde i databasen.

For tiden benytter en administrator et database-forvaltningsverktøy for å inspisere tabeller og attributtene i en informasjonsmengde med tanke på å konfigurere en indeks. Da attributtnavn ofte er mindre enn lesbare, kan en forbetraktning av dataene benyttes til å lette administratorens oppgave med å velge attributter. Denne prosessen kalles skjemaoppdagelse.

I store bedriftssystemer kan det være titusener av tabeller, hver med hundrevis av attributter. Følgelig kan skjemaoppdagelse være en kompleks og tidkrevende prosess.

Det er således en hovedhensikt med den foreliggende oppfinnelse å skaffe søkedrevet skjemaoppdagelse som unngår og eliminerer de ovennevnte ulemper ved nåværende metoder for skjemaoppdagelse.

En annen hensikt med den foreliggende oppfinnelse er å muliggjøre spesifikasjon av informasjonsgjenfinning på basis av skjemaoppdagelsen.

Nok en annen hensikt med den foreliggende oppfinnelse er å forbedre og forenkle resultatnavigasjon med informasjon fra skjemaoppdagelsen.

Endelig er det også en hensikt med den foreliggende oppfinnelse å forbedre søkeapplikasjoner ved å utplassere midler utledet av en skjemaoppdagelsesprosess.

De ovennevnte hensikter så vel som ytterligere trekk og fordeler realiseres med en fremgangsmåte i henhold til den foreliggende oppfinnelse som er kjennetegnet ved å å ekstrahere skjemaveiene forbundet med tilsvarende dokumenter eller innhold, idet skjemaveiene hver omfatter ett eller flere distinkte elementer valgt blant en tjeneradresse, et databasenavn, et dokument eller et attributtnavn,
å beregne sammendragsinformasjon for de ekstraherte skjemaveier, og
å benytte den beregnede sammendragsinformasjon til å danne en indeks basert på søkedrevet skjemaoppdagelse (SDSD-indeks).

I en fordelaktig utførelse av den foreliggende oppfinnelse konstrueres en spesifikasjon for informasjonsgjenfinning på basis av den beregnede sammendragsinformasjon til å.

I en annen fordelaktig utførelse av den foreliggende oppfinnelse benyttes den beregnede sammendragsinformasjon som et hjelpemiddel for resultatnavigasjon i systemet for søking og gjenfinning av informasjon.

I nok en annen fordelaktig utførelse av den foreliggende oppfinnelse samles aksessinformasjon forbundet med en utført søkeapplikasjon ved hjelp av den beregnede sammendragsinformasjon, én eller flere aksessjablonger etableres på basis av den innsamlede aksessinformasjon, og nevnte én eller flere aksessjablonger anbringes i systemet for søking og gjenfinning av informasjon for å forbedre fremtidige søkeapplikasjoner i systemet.

Ytterligere trekk og fordeler vil fremgå av de resterende vedføyde uselvstendige krav.

Den foreliggende oppfinnelse vil forstås bedre når den følgende detaljerte beskrivelse av visse utførelser av den foreliggende oppfinnelse leses i samband med den vedføyde tegning, på hvilket

fig. 1 illustrerer et blokkdiagram av en forenklet søkemotorarkitektur,

fig. 2 et meget minimalt eksempel på tabeller med verdier,

fig. 3 hvordan attributtverdier fra fig. 2 kan representeres i en indeks for å støtte søkedrevet skjemaoppdagelse,

fig. 4 et eksempel på en resultatmengde som omfatter skjemaveier og virkelige verdier fra et eksemplarisk søk,

fig. 5 en forenklet fremleggelse av resultatmengden på fig. 4, hvor de virkelige verdier ikke er vist, og duplikatverdier for skjemaveier fjernet,

fig. 6 hvordan forskjellige tabeller kan sammenføyes, og

fig. 7 fremleggelse av resultater som innbefatter forekomstfrekvenser i skjemaveien.

Før det gås over til drøftelse av foretrukkede utførelser, skal den generelle bakgrunn for den foreliggende oppfinnelse kort beskrives. Som et eksempel kan det forestilles at en administrator for et tids- og kostnadssystem ønsker å

generere en liste over hvilke av hans ressurser som ble tilordnet til eller arbeidet med hvilke prosjekter. Med den nåværende teknologi ville skjemaoppdagelse være en navigasjonsprosess hvor det først må velges en database, deretter en tabell innenfor databasen, og påfølgende dette må attributtnavn og -verdier innenfor tabellen gjennomgås. Navnene vil ofte ikke være intuitive, og det vil være mange å velge blant, så dette er en tidkrevende og frustrerende prosess.

Med søkedrevet skjemaoppdagelse forandrer prosessen seg fundamentalt. Det kan forestilles en database lik den vist på fig. 2. Administratoren begynner med å spesifisere et eksempel på ett av feltene som behøves i resultatet: "Jeg vet ikke hvor denne størrelsen er representert, men jeg vet at jeg har en slik størrelse som har navnet 'John'". Fig. 2 kan tas som en illustrasjon på et meget minimalt eksempel på tabellene 201 ResourceT, 202 CustomerT og 203 ProjectT med verdier og viser i tabellen 204 "ResProjV" hvordan tabeller kan sammenføyes. Tabellen 205 "PP_View" viser hvordan brukeren vil oppfatte data fra denne relasjonen. Verdien "John Smith" har en skjemavei "DB_X.CustomerT.RName". Skjemaveien "DB_X.ResourceT.Person" adresserer verdiene "John" og "Peter", og viser hvordan attributtverdier fra fig. 2 kan representeres i en indeks SDSD for å støtte søkedrevet skjemaoppdagelse som eksemplifiserer en resultatmengde av skjemaveier og naturlige verdier som funnet i en søkeapplikasjon. Denne indeksen er vist på fig. 3 og fremlegger en fullstendig avbildning av slike verdier som gitt ved tabellen 201, 202, og 203 på fig. 2. Basert på dette vil skjemaoppdagelsessystemer tilbake rapportere de forskjellige triplene database-tabell-attributt som har minst én verdi som stemmer overens med dette navnet som gjengitt i listen på fig. 4 og vist forenklet på fig. 5 ved å fremlegge en resultatnavigasjon i stedet for fullstendige resultater. På denne basis kan nå administratoren velge hvilken verdi som er den korrekte.

Denne prosessen gjentas for hvert av attributtnavnene som ønskes i resultatmengden. Etter hvert som nye attributtnavn adderes til denne mengden, ser systemet på måter for sammenføyning over de navngitte attributter eller andre attributter i de samme dokumenter for å skaffe en enhetlig dokumentdefinisjon som inneholder alle attributtnavn.

Basert på denne sammenføyningen kan systemet også tilby andre attributter som foreligger i disse sammenføyde tabeller og som kunne være kandidater for tilføyning til resultatmengden.

For strukturerte informasjonskilder inneholder et dokument en mengde av attributter. Hvert av disse attributtene har et navn som er felles over alle dokumenter. For hvert dokument har hvert attributt også en verdi som kan være eller ikke behøver å være entydig for hvert dokument, og som kan være null (ikke innstilt), inneholde en verdi eller inneholde en mengde av verdier. Foretrukket blir bare enkeltverdier benyttet for entydige attributter til dokumentene i magasinet.

Mengden av attributter for hver dokumentmengde betegnes som skjemaet for dokumentmengden eller tabellen.

En mengde av dokumenter eller poster kan betegnes som en dokumentmengde. Hvis dokumentmengden inneholder alle dokumenter med det samme skjema for en informasjonsmengde, blir mengden ofte implementert som en databasetabell.

Søking er prosessen for å finne et dokument basert på en partiell spesifikasjon av én eller flere av dets attributter. For å forbedre ytelsen til en søkeapplikasjon dannes det ofte en indeks basert på én eller flere innholdskilder. Prosessen for å fylle en indeks med informasjon blir ofte kalt innholdsfangst, og enhver analyse av dataene betegnes som en innholdsforfining.

Med hensyn til den egentlige søkeapplikasjon, dvs. med hvilken informasjonen blir gjenfunnet i databasen ved å benytte et søkespørsmål på den søkbare database og søkeapplikasjonen behandles av en søkemotor som f.eks. drøftet i innledningen av søknaden, kan søkeresultatet gjenfinnes på basis av en identisk eller eksakt overensstemmelse eller en partiell eller tilnærmet overensstemmelse eller ved å innbefattes i en begrepsklasse for én eller flere attributtverdier. I det siste tilfelle kan en begrepsklasse spesifiseres som en person eller organisasjon. I tillegg kan søkespørsmålet benyttes med en lingvistisk normalisering for å forbedre gjenkall i søkeresultatet, idet gjenkall er et mål på de returnerte dokumenter i søkeresultatet. Hvis lingvistisk normalisering benyttes på et søkespørsmål, kan dette foretrukket gjøres ved hjelp av lemmatisering, vanlig stavekontroll, fonetisk

overensstemmelse, synonymer og homeosemier, idet de sistnevnte betegner nærsynonymer. Alle disse foretrukkede tiltak i forbindelse med søkeapplikasjon kan betraktes som velkjente for fagfolk innenfor området søking og gjenfinning av informasjon.

Strukturerte kilder inneholder typisk en mengde av databasetabeller, og noen av disse kan det være nødvendig å sammenføye for å frembringe søkbare objekter. Prosessen med å velge slike tabeller, å konfigurere de verdier som det skal sammenføyes over og å velge hvilke dokumenter som skal mates til indeksen, kalles indekskonfigurering. For meningsfylt å konfigurere en indeks, må en administrator forstå skjemaet til datatabellene.

For nærværende benytter en administrator et databaseadministrasjonsverktøy for å inspisere tabeller og attributter i en informasjonsmengde med tanke på å konfigurere en indeks. Da attributtnavn som nevnt, ofte er mindre enn lesbare, skaffes en forbetraktning av data for å lette administratorens oppgave ved valget av attributter. Denne prosessen kalles skjemaoppdagelse.

Skjemaveien til et attributt er en eksakt beskrivelse av hvor et attributt kan finnes. Dette vil i en database typisk omfatte a) tjeneren hvor databasen befinner seg, b) navnet på databasen, c) navnet på tabellen, og d) navnet på attributtet, eller i en alternativ notasjon "server.db.table.attribute".

Spesielt vil fremgangsmåten i henhold til den foreliggende oppfinnelse muliggjøre bruk av søkedrevet skjemaoppdagelse for å finne skjemaet til en SQL-database. I nåværende databasesystemer omfatter skjemaoppdagelse bruk av et databaseforvaltningssystem til manuelt å inspisere hver tabell eller en undermengde av tabeller valgt etter navn for å se om verdiene er de man behøver. I store bedriftssystemer kan det være titusener av tabeller, hver med hundrevis av attributter. Følgelig kan som angitt ovenfor skjemaoppdagelse være en kompleks og tidkrevende prosess. I slike systemer bestemmer vanligvis i tillegg navnekonvensjoner hvilke navn som kan benyttes for alle størrelser, slik at navnene typisk ikke er intuitive for en menneskelig bruker. Ved den foreliggende oppfinnelse vil brukeren starte med eksempler som er kjent å foreligge i dataene, kjøre søkespørsmål basert på disse, og søkesystemet vil tilby kandidatattributter som brukeren kan inspisere.

Fremgangsmåten i henhold til den foreliggende oppfinnelse benyttes til å oppdage strukturen til data lagret i XML. I et nåværende XML-basert system

vil en bruker manuelt kjøre XQuery-spørsmål eller benytte en XQuery-basert leser for å inspisere innholdet i systemet. Den foreliggende oppfinnelse vil indeksere den underliggende informasjon og la brukeren kjøre en søking, noe som resulterer i kandidatsteder for denne ønskede informasjon.

I en foretrukket utførelse av den foreliggende oppfinnelse kan en spesifikasjon av informasjonsgjenfinningen konstrueres. Hvordan dette gjøres, er vist på fig. 6. Et attributt velges fra tabellen 601 "ResourceT" og et attributt fra tabellen 601 "ProjectT". Nå kan det bestemmes fra databaseskjemaet at disse tabellene kan sammenføyes over tabellen 601 "ResProjV", og basert på dette forhold blir informasjonsgjenfinnings-spesifikasjonen 604 generert som vist. Slik det fremgår av fig. 6, ses det at i dette eksempel tar spesifikasjonen 604 for informasjonsgjenfinning form av et SQL-utsagn.

I denne utførelse kan det søkedrevne skjemaoppdagelse benyttes til å foreta flytting av programvaresystemer for bedrifter eller foretak. Med kjent teknologi kan en bedrift som ønsker å oppgradere et bedriftsprogramvaresystem måtte gå gjennom en manuell prosess hvor strukturen til kandidatsystemet inspiseres for å avdekke tilpasninger og bruksmønstre. Dette må da gjenspeiles i det nye system. For store bedrifter som flytter fra en leverandør av Enterprise Resource Planning (ERP) til en annen, er denne oppgaven kjent å kreve investeringer på mange millioner dollar og ta flere år. Skjemaoppdagelse er en vesentlig del av denne kostnaden. Hele prosessen er basert på en god forståelse av det virkelige underliggende skjema og kunne gjøres mye mer effektiv ved søkedrevet skjemaoppdagelse.

En spesifikasjon for informasjonsgjenfinning som generert i den første utførelse av den foreliggende oppfinnelse, kan også benyttes til å redusere kostnaden ved å generere rapporter i et bedriftsprogramvaresystem. Med den nåværende teknologi er en manuell prosess for å velge tabeller som skal benyttes som basis for rapporter tidkrevende og tilbøyelig til å medføre feil. Med fremgangsmåten i henhold til den foreliggende oppfinnelse ville seleksjonsprosessen være eksempeldrevet. Ta et eksempel hvor en bruker behøver å generere en salgsrapport til kunder. Med nåværende teknologi ville brukeren starte med å se på tabellnavn eller betrakte navnet, og sannsynligvis lete etter tabellnavn som inneholdt begreper som "sale" eller "customer".

Hvis en slik tabell finnes, vil brukeren se på verdiene for å sjekke om det er sannsynlig at den funne informasjon er den korrekte. Denne prosessen blir usedvanlig tungvint i systemer hvor navngivningskonvensjoner ikke er intuitive, da brukeren på forhånd må se alle tabellene i systemet. Prosessen er også utsatt for feil, fordi det er mange tilfeller hvor tilsvarende data holdes i flere tabeller og benytte for noe forskjellige formål. Et system basert på den foreliggende oppfinnelse ville be brukeren om et eksempel på en slik kunde, f.eks. "ACME". Et søk ville deretter utføres, og resultatet kunne være at "dette navnet forekommer i følgende tabeller: current_customers, former_employers, and marketing_partners". Av dette utvalget ville brukeren uten videre vite hvem rapporten skulle baseres på. Hvis de samme tabellene var skjult under navnet XCC_1543, XCB_2063, og XAA_M15 i et system som også omfatter 20 000 tabeller til, ville evnen til å fokusere på slik liten undermengde være vesentlig med tanke på å få jobben gjort.

Fremgangsmåten i henhold til den foreliggende oppfinnelse skaffer en forenkling av prosessen med å velge en undermengde av tabeller og attributter for å gjøre dem søkbare i en søkeindeks. Med den nåværende teknologi må skjemaet enten være *a priori* kjent eller den samme tungvinte, manuelle oppdagelsesprosess utføres. Med søkedrevet skjemaoppdagelse vil et kandidatundermengde typisk returneres i form av nedboringer ("drilldowns") som tillater brukeren å velge de ønskede attributter.

Når resultater fremlegges, er den mest vanlige representasjon en resultatliste. Dette blir tungvint hvor mange resultater er tilgjengelige, da resultatene som virkelig behøves kan forekomme lenger ned i listen enn et stort antall andre treff. Som eksempel kan man forestille seg at den foreliggende oppfinnelse benyttes til å søke etter verdien "John", og at tabellen inneholder 1000 referanser som innbefatter "John", i tabell A og bare én i tabell B. En resultatpresentasjon uten navigasjon ville kreve at brukeren gikk gjennom alle treffene fra tabell A før treffene fra tabell B ble funnet. Dette er vist i listene 701 på fig. 7. Knappen "NEXT" lar brukeren se den neste undermengde.

Nok en annen foretrukket utførelse av den foreliggende oppfinnelse fremlegger resultater ikke som en liste, men som en resultatnavigasjon. Kort sagt blir resultatnavigasjonen fremlagt som en forbundet liste as skjemaveier. Forbedringen her vil skaffe en gruppering på tabeller og tillate brukeren å

velge "A" eller "B" for å navigere til det eneste dokument som passer med denne spesifikasjonen med bruk av skjemavei 702, vist på fig. 7. En ytterligere forbedring av dette teller resultatene for å vise brukeren at antallet tilsvarende resultater for hvert navigasjonsvalg, som vist ved skjemaveien 703, og tillater dermed at frekvensinformasjon om forekomst innbefattes i listen av skjemaveier.

Ytterligere en annen foretrukket utførelse av den foreliggende oppfinnelse vil tilby en sterkt redusert innsats og også å redusere startiden for å gjøre store magasiner søkbare. Uten indeksering omfatter søking i store magasiner typisk en skanning av data, noe som er en meget tidkrevende prosess. Selv med den nåværende teknologi, blir dokumenter som skal gjøres søkbare, typisk avnormalisert for å kombinere verdier som det sammen skal søkes etter. Med fremgangsmåten i henhold til den foreliggende oppfinnelse og et søkesystem som støtter sammenføyning, ville først alle primærverdier indekseres, dvs. ikke gjentatte verdier i individuelle attributter i datavarehuset. Deretter kunne en kompleks søking utføres mot hvert attributt, og resultatene sammenføyes for å finne det virkelige resultat.

Fremgangsmåten i henhold til den foreliggende oppfinnelse vil deretter bli benyttet til å klarlegge kombinasjonen av attributter benyttet i virkelige søk. Denne informasjon kunne deretter brukes til å danne en fysisk indeks for de kombinasjoner av attributter som det faktisk søkes etter, og således benytte et iakttatt søkemønster så å si som en sjablong for aksessoptimering. Med dette system installert, vil brukeren ha muligheten av å eksekvere søk, om enn langsomt, meget tidlig i prosessen, f.eks. i løpet av noen dager istedenfor over kanskje et år. Over tid vil deretter aktuelle søkemønstre kunne benyttes som basis for å danne en indekskonfigurasjon optimert mot disse søkemønstre og dermed forbedre søkeytelsen.

PATENTKRAV

1. Fremgangsmåte til å danne grensesnitt mellom applikasjoner for søking, analyse og rapportering i et system for søking og gjenfinning av informasjon med dokument- eller innholdsmagasiner som inneholder komplekse, strukturerte dokumenter eller innhold, hvor magasinet er søkbart og omfatter skjemaveier for dokument- og innholdsattributter, og hvor fremgangsmåten omfatter
trinn for å benytte et søkespørsmål for én eller flere attributtverdier på en indeks av attributtverdier, og gjenfinne en resultatmengde av dokumenter eller innhold som tilsvarer nevnte én eller flere attributtverdier, og hvor fremgangsmåten er
k a r a k t e r i s e r t v e d
å ekstrahere skjemaveiene forbundet med tilsvarende dokumenter eller innhold, idet skjemaveiene hver omfatter ett eller flere distinkte elementer valgt blant en tjeneradresse, et databasenavn, et dokument eller et attributtnavn,
å beregne sammendragsinformasjon for de ekstraherte skjemaveier, og
å benytte den beregnede sammendragsinformasjon til å danne en indeks basert på søkedrevet skjemaoppdagelse (SDSD-indeks).

2. Fremgangsmåte i henhold til krav 1,
k a r a k t e r i s e r t v e d å beholde bare enkeltverdier for entydige attributter til dokumentene i magasinet.

3. Fremgangsmåte i henhold til krav 1,
k a r a k t e r i s e r t v e d å gjenfinne søkeresultatet på basis av én en identisk eller eksakt overensstemmelse, en partiell eller tilnærmet overensstemmelse eller ved at det er innbefattet i en begrepsklasse for nevnte én eller flere attributtverdier.

4. Fremgangsmåte i henhold til krav 3,
k a r a k t e r i s e r t v e d å spesifisere en begrepsklasse som en person eller en organisasjon.

5. Fremgangsmåte i henhold til krav 1,
k a r a k t e r i s e r t v e d å benytte søkespørsmålet med lingvistisk normalisering for å forbedre gjenkall i søkeresultatet.

6. Fremgangsmåte i henhold til krav 5,
k a r a k t e r i s e r t v e d å utføre lingvistisk normalisering med én eller flere blant lemmatisering, stavelseskontroll, fonetisk overensstemmelse, synonymer eller homeosemier.

7. Fremgangsmåte i henhold til krav 1,
k a r a k t e r i s e r t v e d å konstruere en spesifikasjon for informasjonsgjenfinning på basis av den beregnede sammendragsinformasjon.

8. Fremgangsmåte i henhold til krav 7,
k a r a k t e r i s e r t v e d å formulere spesifikasjon for informasjonsgjenfinningen som et SQL- eller XQuery-utsagn.

9. Fremgangsmåte i henhold til krav 8,
k a r a k t e r i s e r t v e d å overføre informasjon fra magasinet til et annet system for gjenfinning og søking av informasjon ved hjelp av et SQL-utsagn.

10. Fremgangsmåte i henhold til krav 9,
k a r a k t e r i s e r t v e d at det annet system for søking og gjenfinning av informasjon er ett blant database, datavarehus, rapporteringssystem, søkemotortjeneste eller et applikasjonsprogrammert grensesnitt.

11. Fremgangsmåte i henhold til krav 1,
k a r a k t e r i s e r t v e d å benytte den beregnede sammendragsinformasjon som et hjelpemiddel for resultatnavigasjon i systemet for søking og gjenfinning av informasjon.

12. Fremgangsmåte i henhold til krav 11,
k a r a k t e r i s e r t v e d å fremlegge resultatnavigasjonen som en liste av forbundne skjemaveier.

13. Fremgangsmåte i henhold til krav 12,
k a r a k t e r i s e r t v e d å innbefatte informasjon om forekomstfrekvens i listen av skjemaveier.

14. Fremgangsmåte i henhold til krav 11,
k a r a k t e r i s e r t v e d å samle aksessinformasjon forbundet med en utført søkeapplikasjon ved hjelp av den beregnede sammendragsinformasjon, å etablere én eller flere aksessjablonger på basis av den samlede aksessinformasjon, og

å anbringe de nevnte én eller flere aksessjablonger i søkesystemet for gjenfinning og informasjon for å forbedre fremtidige søkeapplikasjoner i systemet.

5

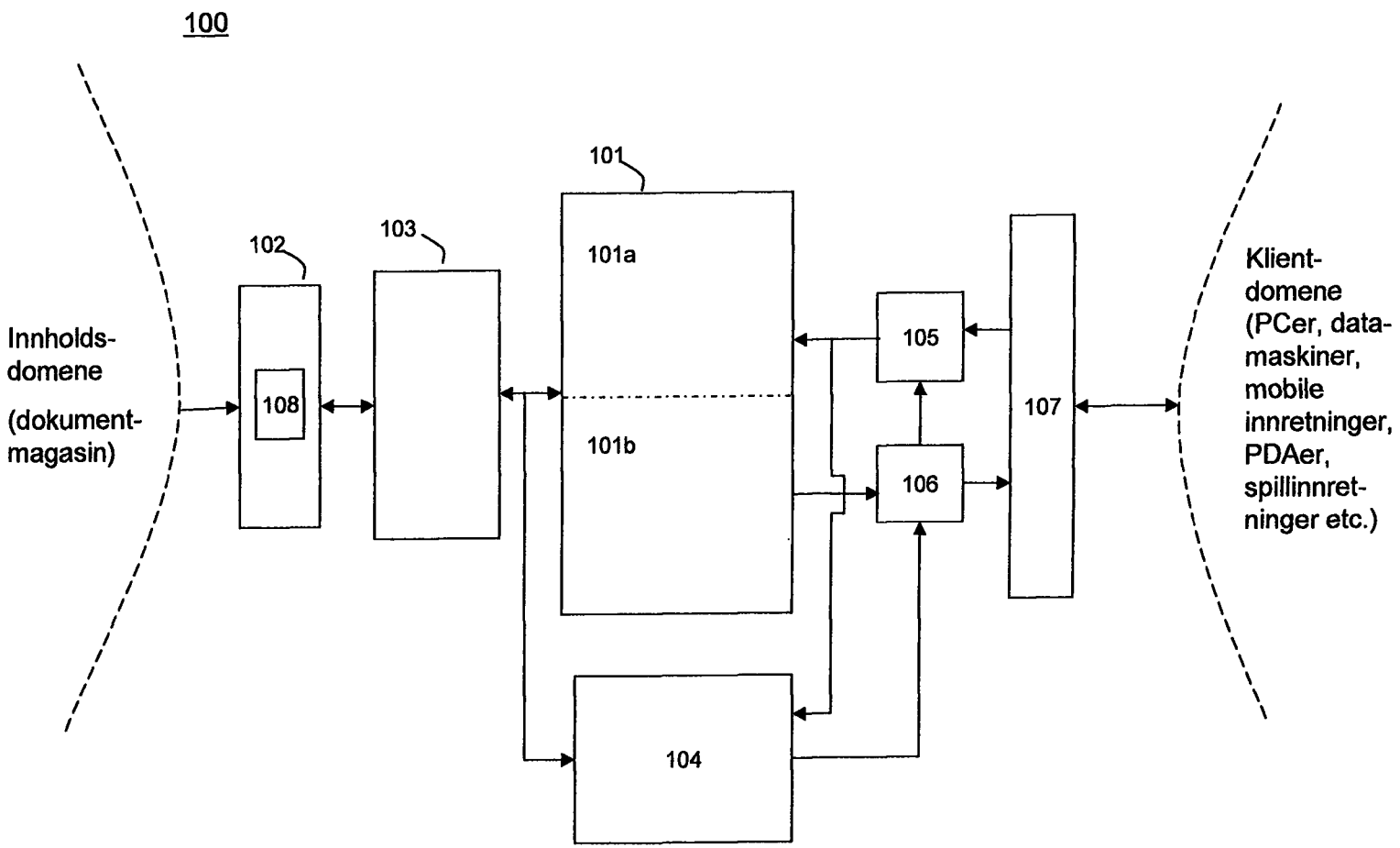

100
Innholds-
domene
(dokument-
magasin)
102
108
103
101
101a
101b
104
105
106
107
Klient-
domene
(PCer, data-
maskiner,
mobile
innretninger,
PDAer,
spillinnret-
ninger etc.)

Fig. 1