

ETDEWEB versus World-Wide-Web (Google /Google Scholar)

En spesifikk database/Web sammenligning

Debbie Cutler¹

Energy Technology Data Exchange (ETDE) Operating Agent Energy DOE/OSTI, Oak Ridge, TN (United States) DOE / OSTI, Oak Ridge, TN (USA)

Publisering dato: juni 2010

Oversatt av Astrid Gudmundseth

Sammendrag: Dette er en studie hvor søkeresultater fra den spesialiserte, vitenskapelige databasen innen energirelatert informasjon, ETDEWEB² blir sammenlignet med søkeresultater fra søkemotorene Google og Google Scholar (GS). Det primære målet med denne studien var å finne ut om ETDEWEB fortsatt gir brukeren søkeresultat som ikke blir funnet av Google eller Google Scholar.

ETDE er et internasjonalt program for å utveksle informasjon, hvor medlemsland og partnere gjennom kostnads- og oppgavedeling bidrar til verdens største database av energirelatert informasjon.

Tidlig i 2010 passerte databasen ETDEWEB 4,3 millioner henvisninger til energilitteratur fra hele verden.

En av styrkene til ETDEWEB er fokuset på vitenskapelig innhold og direkte tilgang til fulltekst for grå litteratur (over 300 000 dokumenter i PDF er tilgjengelig for visning fra ETDEs område og over en million lenker til dokumenter hos forskningsorganisasjoner og store utgivere globalt). Google og GS er godt kjent for at de søker i den store bredden av informasjon. Google bringer nyheter, fakta- og meningsrelatert informasjon og GS vektlegger faglig innhold innen flere fagfelt.

Analysen sammenligner resultatene av 15 energirelaterte søk som er utført på alle tre systemene og hvor det er brukt identiske ord / fraser. En rekke emner ble valgt, selv om disse hovedsakelig var innenfor fornybar energi på grunn av bred, internasjonal interesse for dette.

Over 40 000 poster fra de tre kildene ble evaluert. Studien konkluderer med at ETDEWEB er en betydelig ressurs for energi eksperter for å oppdage relevant informasjon. For de 15 temaene i denne studien, har ETDEWEB gitt brukeren unike resultater som ikke vises av Google eller GS i 86,7 % av tilfellene.

Mye ble lært i studien. Observasjoner om styrkene i hvert system og andre faktorer som påvirker søkeresultatene deles sammen med bakgrunnsinformasjon og tabeller som oppsummerer resultatene. Hvis en bruker vet den eksakte tittel på et dokument, er alle tre systemene nyttige for å finne kilde til dokumentet. Hvis brukeren er ute etter å finne relevante dokument innen et bestemt emne, vil hver av de tre systemene gi en betydelig mengde data, men ganske ulike i fokus. Google er et mye brukt og nyttig verktøy for å finne betydelige "ikke-spesialist" informasjon, og GS hjelper brukeren med fokus på vitenskapelige disipliner. Men hvis en brukers interesser er vitenskapelige og energispesifikke, har fortsatt ETDEWEB mer energiforskning og utvikling (FoU) og en betydelig merverdi i kunnskap om ny teknologi.

¹ Forfatteren setter pris på tilbakemeldingene fra ETDEs medlemsorganisasjoner gjennom hele prosessen med studien, fra innledende resultater til offentliggjøringen av rapporten.

² ETDEWEB – Energy Technology Data Exchange – World Energy Base

ETDEWEB versus World-Wide-Web

Det verdensomspennende nettet har dramatisk endret måten de fleste bedrifter og enkeltpersoner søker etter informasjon på. Internett har blitt en integrert del av dagliglivet til en stadig voksende del av verdens befolkning både på arbeidsplassen og i hjemmet. Google og andre søkemotorer har forsterket en oppfatning av at "alt finnes der ute" ved å gi mange treff for nesten alle søk. Bibliotekbudsjett blir gransket mer nøye enn noen gang. Arbeidstakerne må finne informasjon på egen hånd, fordi "alt lett kan finnes weben". Mengden er ikke noe problem når det gjelder søkeresultater, men hva med andre aspekter? Hvem kontrollerer kvaliteten og fullstendigheten av informasjon og hva søkemotorene "finner"?

Hvem er ansvarlig for å få informasjon ut på nettet? Hvor mye tid bruker enkeltpersoner til individuell filtrering i havet av informasjon? Tror de at det de ser, er det beste eller den mest pålitelige tilgjengelige informasjonen? Eller er spesialiserte databaser og informasjonsportaler mer nyttig for brukerne? Og hvis så, gir de verdiøkninger?

Disse spørsmålene er av spesiell interesse i det vitenskapelige samfunn, der kvaliteten på informasjonsdeling er avgjørende for utviklingen av vitenskapen. For verdens største vitenskapelige database på energiområdet ETDEWEB har en sammenlignende studie blitt gjennomført.

I studien blir søkeresultatene fra ETDEWEB sammenlignet med resultater fra to av de mest populære søkemotorene, Google og GS. Det primære målet var å finne ut om ETDEWEB fortsatt gir brukeren søkeresultater som ikke blir funnet av søkemotorer og med det illustrerer sin spesielle verdi. Studien konkluderer at selv med overflod av informasjon i den verdensomspennende weben, fortsetter ETDEWEB å være en betydelig, nisjeressurs for energiexperteer og andre for å oppdage relevant energiinformasjon. Analysen viser at ETDEWEB gir søkeresultater som er unike i 86,7 % av tilfellene (i gjennomsnitt) i forhold til resultatene for samme søk utført på Google eller GS. Flere erfaringer enn statistikk ble høstet i studien. Rapporten gir bakgrunnsinformasjon om studien og en oppsummering av resultatene. Observasjoner angående styrken i hvert system og faktorer som påvirker søkeresultatene blir også delt.

Bakgrunn

Ideen til ETDEWEB- studien var tidligere analyser (2009) som sammenlignet søkeresultater fra WorldWideScience (WWS)³ med Google og GS. Siden WWS var et nytt informasjonssystem, ønsket sponsorene å finne ut hvordan innholdet i deres system kom ut i forhold til hva som ellers kan finnes på nettet. På samme måte har ETDEWEBs sponsorer også vært opptatt av å vite hvordan ETDE- databasen kommer ut sammenlignet med Google / GS. Derfor ble det iverksatt en egen studie hvor man fulgte en lignende metode.

ETDEWEBs sponsor er ETDE, Energy Technology Data Exchange, en internasjonal avtale for å utveksle informasjon dannet i 1987 under rammen av Det internasjonale energibyrået (IEA).

ETDEs medlemsland bidrar med finansiering og både medlemmer og partnere bidrar med poster / dokumenter som representerer verdensomspennende forskning, vitenskap, teknologi og FoU innen energi inkludert politikk, miljømessige og økonomiske aspekter. ETDEs oppgave er å gi regjeringer, industri og forskningsmiljøer i medlemslandene tilgang til innsamlet informasjon og øke spredningen til utviklingsland. Over 110 land har fri tilgang til ETDEWEB, og ETDE ønsker velkommen interesse fra land som ikke allerede har tilgang. ETDEs mål er oppnådd gjennom etableringen av Energy Database.

Webversjonen, ETDEWEB (ETDE World Energy Base) (<http://www.etde.org/etdeweb>) ble brukt i sammenlikningen.

Studien ble utført av ETDEs Operating Agent, Department of Energy's Office of Scientific and Technical Information som vedlikeholder ETDEWEB.

³ en informasjonsportal

ETDEWEB er kjent som den største databasen av energirelatert informasjon i verden. Per tidlig i 2010 har databasen over 4,3 millioner henvisninger til verdensomspennende energilitteratur. En av styrker er dens fokus på vitenskapelig innhold, med fulltekst for den grå litteraturen (over 300,000 dokumenter i PDF og daglig økende) tilgjengelig for visning direkte fra ETDEs nettside. Over en million ekstra poster inneholder lenker til dokumenter hos forskningsorganisasjoner og store utgivere globalt. ETDEWEBs underliggende databasestruktur gir mulighet for enkle eller avansert søk og andre funksjoner som hjelper brukerne i deres avgrensning av resultater. Informasjonen i ETDEWEB forventes å være av vitenskapelig karakter og er filtrert av medlemslandene før innlegging. Kilder inkluderer rapporter fra store forskningsorganisasjoner, tekniske konferanser, fagfelleurderte tidsskrifter og mye mer. En bruker som søker ETDEWEB er ikke nødt til å bla seg gjennom pressemeldinger, produktkampanjer, annonser, og sosiale mediers nettsted for å se forskningsresultater. Dokumenter som er på andre språk har engelsk tittel og sammendrag inkludert i databasen, dette for å hjelpe brukere å finne relevant faglig innhold. Det hjelper også med å avgjøre om det er verdt å oversette hele teksten. Emnemessig indeksering ved hjelp av kontrollert vokabular er lagt til hver referanse og er også til hjelp for mer presise søk. Et nytt søkeverktøy for å innsnevre søk - emneklynger (subject clustering) ble tilgjengelig i ETDEWEB i februar 2010. Dette verktøyet utnytter emneindekseringen.

Google er et bemerkelsesverdig produkt som har forandret brukeropplevelsen, og google er også tatt i vanlig bruk som verb. Dens styrke ligger i stor mengde av informasjon raskt tilgjengelig og algoritmer for rangering, med forbløffende evne til å levere informasjon i samsvar med brukerens interesser. Et typisk søk returnerer generelt millioner av "treff", hvor de 10 første vises til brukeren, og resten er tilgjengelig i trinn på ti. Google prøver virkelig å indeksere "alt" den kan få tilgang til. Opplysninger som finnes i Google inkluderer kilder med vitenskapelig innhold fra "overflate" web, men det synes å være større fokus på nyheter, forretningsinformasjon og produkter, blogger, reklamemateriell, anmeldelser, informasjon / referansenettsteder som Wikipedia, og store utsalgssteder som Amazon. Google Scholar, som navnet tilsier, fokuserer vanligvis på ressurser som er antatt å være av interesse for akademisk- og forskningsmiljøer. Viktige tidsskriftforlag, informasjonsforeninger (societies) og mange vitenskapelige databaser tidligere regnet som en del av "the deep web", er gjort lettere tilgjengelig via denne spesialiserte søkemotoren.

Rammene for hvert søkesystem er også nyttig å forstå. Søkemotorer som Google / GS har vanligvis ikke kilden til innholdet, selv om de har mange partnere. Søkemotorene har "bots" eller "crawlers" som oppsøker nettstedet og bygger indekser som hjelper med rangering og i å gjenkjenne lignende søk. Med rangeringen bidras / kontrolleres hva brukeren ser som søkeresultat. For store deler av opplysninger på nettet, må en person eller et foretak gjøre en bevisst innsats for å sikre at informasjonen de produserer, er satt i et godkjent format og er synlig for "bots" slik at søkemotorene vet hvor og hva som skal indekseres. Spesialistselskapers påstander om å være i stand til å øke synligheten av bedriftens informasjon på nettet, er voldsomme, men det er ikke tvil om at har de noen tips for å gjøre nettopp det. Realiteten er at ikke alle søkemotorer søker de samme kildene, heller ikke at de gjør det på samme måte. Hyppighet og metode for indeksering varierer betydelig fra en søkemotor til en annen og fra en lokalitet til en annen. Fra nesten øyeblikkelig indeksering for nyheter og værmeldinger, til en mye tregere og/eller bare delvis /utvalgt indeksering av store samlinger – dette også selv om de er gjort kjent for "bots" (mer om dette aspektet er tatt opp senere i rapporten). Det samme søket vil derfor gi forskjellige resultater fra en søkemotor til en annen. Ikke bare på grunn av innhold, men også på grunn av rangeringen av svarene som vises til brukeren. Siden rangeringsalgoritmer generelt er beskyttet, er det ukjent hvor mye vekt det er gitt til aktualitet av informasjonen, nettstedets popularitet, datapålitelighet, sponsorer og / eller forholdet til annonsørene, og andre faktorer. Men tross disse variablene; brukerne er generelt svært fornøyd med resultatene fra Google spesielt, og stiller sjelden spørsmål om det er noen opplysninger som mangler siden de allerede får mer informasjon enn de kan overkomme.

Andre observerte detaljer i studien var at Google og GS håndterte flertallsformer bedre (de finner både cell og cells, for eksempel der ETDEWEB for tiden ikke finner det uten at brukeren spesifikt ber om det), og de har noen grad av semantisk-baserte hjelpemidler for søk som hjelper resultatene. Det er antatt at Google og GS også finner resultater basert på hele teksten i dokumenter, når de er tilgjengelige, mens ETDEWEB søkene som ble gjort for denne studien, er gjort i Easy Search, som ikke inkluderer hele teksten. (Merk: ETDEWEB har fulltekstsøk tilgjengelig, og det ville betydd at enda flere poster ble funnet).

Analysedetaljer og resultater

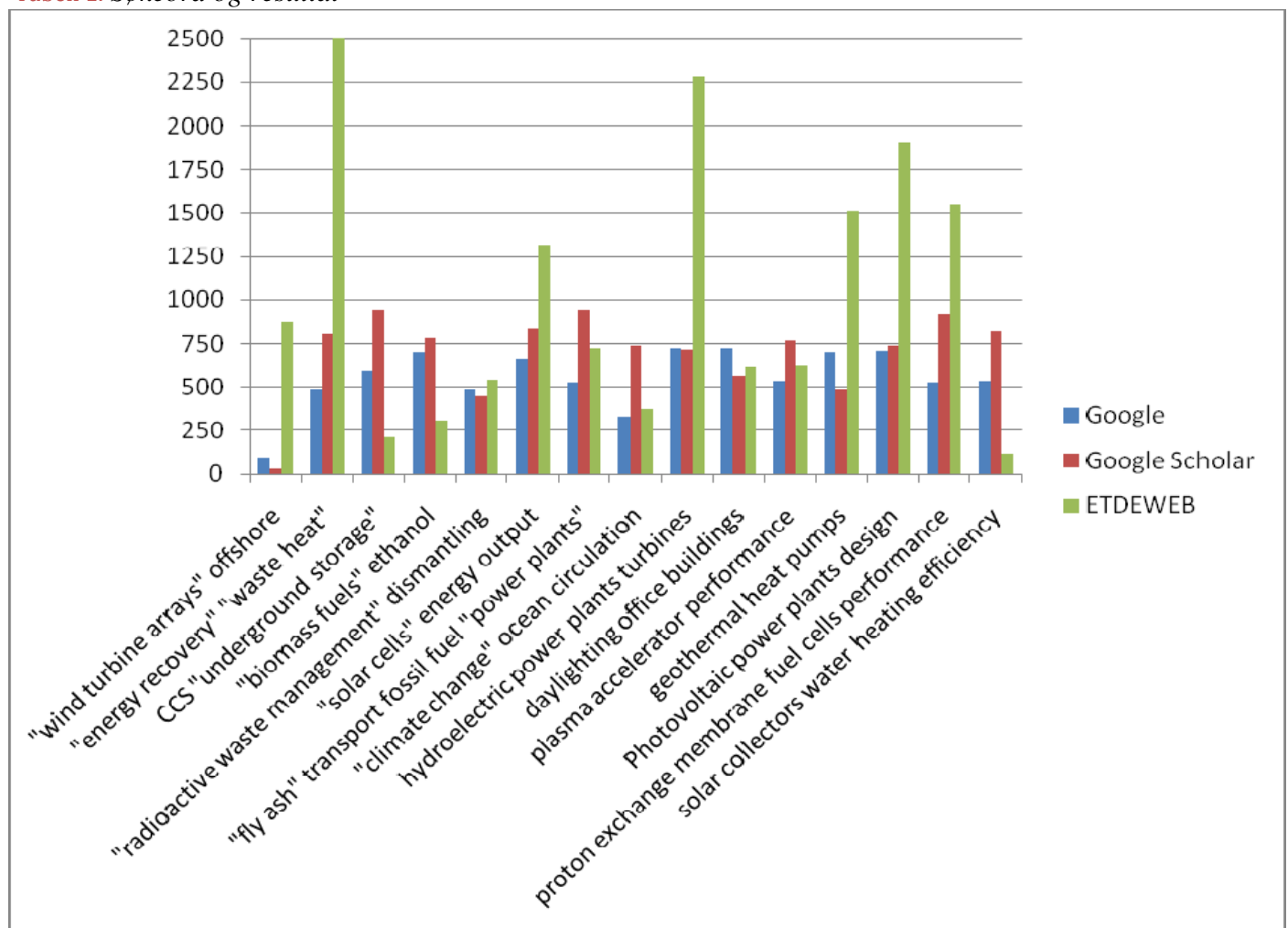
Analysen sammenlignet resultatene av 15 energirelaterte søk som er utført på ETDEWEB, Google og GS og bruker identiske ord. En rekke emner ble valgt, selv om det meste er innen fornybar energi på grunn av høy aktualitet for mange av ETDEs brukere. De termer som ble brukt var antatt å være spesifikke nok til å holde antall poster nede på et håndterbart nivå i stedet for å resultere i stort volum. Resultatene fra hvert søk ble tatt vare på som 45 datasett med titler og kildeinformasjon som så ble brukt for sammenlikninger.

Over 40 000 poster var en del av studiet.

Komprimerte sammenlikninger var et interessant fenomen observert i WWS studien og det samme var det i ETDEWEB studien. Selv om den opprinnelige resultatmengden oppgitt av Google og GS (for eksempel "Resultater 1-10 av ca 658,000") er veldig imponerende, viser det seg i virkeligheten at det maksimale antall referanser, som brukeren faktisk kan bla seg gjennom og se, aldri overstiger 1000. Videre, som brukeren blar gjennom side etter side, avgrensner Google og GS disse mengdene (tilsynelatende eliminer dubletter eller svært like resultater), og det endelige tall vil nesten alltid være lavere enn de mengdene som opprinnelig ble oppgitt. Siden mange brukere vanligvis ikke engang ser utover de første sidene med resultater, er dette mønsteret ikke så tydelig for den vanlige bruker, eller kanskje ikke så viktig. Men det betød at det faktiske antallet poster tilgjengelig for sammenlikning i studien alltid var begrenset til mindre enn 1000 for Google og GS resultater. Når det var datasett, ble det observert ytterligere duplisering innenfor hvert resultatsett. Disse duplikater ble eliminert før den endelige sammenlikningen ble gjort. Tabell 1 viser søkeordene brukt til søk og antall poster funnet i av hvert av søkesystemene når alle duplikater er fjernet.

Tittelen på artikkelen eller dokumentet var fellesnevneren på tvers av systemene og det viktigste feltet brukt til sammenlikninger. Analysen så på kildefeltet i tillegg til andre faktorer, men ikke for direkte samsvar. Nærmere gransking ble foretatt for undergrupper av ETDEWEB poster (for eksempel fagfelleverdert tidsskrift) som var forventet å være til stede i søkeresultatene fra Google og / eller GS, for å sikre at sammenlikningene var korrekte.

Tabell 1. Søkeord og resultat



Som det fremgår av tabell 1, hadde ETDEWEB i gjennomsnitt 1403 poster per søk mens Google hadde 556 og GS 704. Høyere gjennomsnittligstall er åpenbart positive for ETDEWEB, men det er også litt misvisende som gjennomsnitt, gitt begrensningene for Google / GS resultater. ETDEWEB hadde ingen øvre grense på sine resultater selv om en grense på et par tusen var antatt.

Når vi sammenlignet treffene fra ETDEWEB /Google og GS, fant vi at kun 57 treff per søk i gjennomsnitt, var sammenfallende. Det var flere sammenfallende treff i GS enn i Google, dette i et forhold på 17:1. Dette forholdet var ikke uventet siden GS legger vekt på et vitenskaplig/ teknisk innhold. Viktige kilder i GS treff er internasjonale tidsskriftsutgivere, tekniske samfunns publikasjoner (society papers), konferanser, akademiske kilder og også poster fra to av OSTI egne databaser, Information Bridge and Energy Citations. Resultatene fra Google var mye mer variert med innhold som spenner fra blogger, wikier, nyhetsartikler, pressemeldinger, næringsliv, myndigheter og forbrukerorienterte nettsteder til patenter, bøker fra utgivere og produktbeskrivelser. Et av resultatene undersøkelsen forventet å se, var at noen av treffene i Google ville være titler hvis opprinnelse var ETDEWEB. ETDE gjorde en betydelig del av sitt databaseinnhold tilgjengelig for Googles "bots" og andre søkeroboter tidlig i 2009, men det var skuffende sjelden å se en ETDEWEB post i Googlelisten. Oppfølgende testing viste at ETDE fortsatt ikke har høy rangering i resultatlisten i de fleste tilfellene når man søker på Google, eller eventuelt at postene er fjernet som en del av søkemotorens utvelgelsesprosess.

En nærmere gjennomgang av robotsøkeprogrammets aktivitet i 2010, anslår at Google indekserer mindre enn 10 % av ETDEWEBs poster som gjøres tilgjengelig for det. Man undersøker nå hvilke tiltak som kan gjøres for å øke indekseringen av ETDE. Hvis ETDE informasjonen virkelig ikke er "der ute" tross alle anstrengelser for å gjøre den tilgjengelig, hvor mange andre samlinger er også bare delvis indeksert?

Det neste steget i undersøkelsen var å sammenligne de ikke-sammenfallende referansene fra den internasjonale tidsskriftutgiveren Elsevier. GS indekserer tidsskrift fra Elsevier.

Denne grundige gjennomgangen førte til ytterligere avgrensninger i treffene. Et tillegg på 129 treff (i gjennomsnitt) ble gjort etter disse avgrensninger. Tabell 2 gir prosentandelen av ETDEWEB resultater som var unikt for samme søk i hvert system etter disse avgrensningene.

Tabell 2. Prosent av dokument funnet i ETDEWEB og som ikke finnes av Google eller GS

Søketermer	Prosent unike for ETDEWEB Resultat
"wind turbine arrays" offshore	90,00 %
"energy recovery" "waste heat"	87,30 %
CCS "underground storage"	87,10 %
"biomass fuels" ethanol	87,80 %
"radioactive waste management" dismantling	85,40 %
"solar cells" energy output	87,30 %
"fly ash" transport fossil fuel "power plants"	77,60 %
"climate change" ocean circulation	85,90 %
hydroelectric power plants turbines	87,80 %
daylighting office buildings	82,30 %
plasma accelerator performance	84,50 %
geothermal heat pumps	86,40 %
photovoltaic power plants design	88,10 %
proton exchange membrane fuel cells performance	86,00 %
solar collectors water heating efficiency	86,10 %
Gjennomsnitt	86,70 %

Det neste steget i undersøkelsen var å se om de referansene som var unike for ETDEWEB, kunne finnes i Google og GS. I mange tilfeller ble de funnet ved å søke på den eksakte tittelen. Men igjen, disse referansene ble ikke funnet av Google og GS i emnesøk brukt i studien. Analysen viser at selv om et dokument er tilgjengelig på weben, betyr det ikke at det vil bli funnet i et Google /GS søk.

Mye filtrering og rangering skjer, og hva brukeren vises, mistenkes å være, i alle fall delvis, basert på populariteten til kilden. Mange av postene har slagside mot visse kilder. Og hvis postene er like, var det typisk "prestisje" kilden som ble vist. Studien konkluderte at hvis en bruker vet en helt bestemt tittel, er alle tre systemene nyttige i å finne en kilde for dokumentet. Men hvis brukeren er ute etter å finne relevante dokumenter om et bestemt emne, som var denne studiens fokus, vil hver av de tre systemene gi en betydelig mengde data, men forskjellige.

For emnene i denne studien, har det vist seg at ETDEWEB bringer brukeren unike resultater som ikke vises av Google eller GS i 86,7 % av tilfellene.

Konklusjon

Google gir brukeren et verdifullt verktøy for å finne en overflod av god informasjon, og GS hjelper brukere med fokus på vitenskapelige disipliner. Men hvis en brukers interesser er vitenskapelige og energispesifikke, holder ETDEWEB fortsatt en sterk posisjon innen energi- forskning og utvikling (FoU) og har en betydelig merverdi i kunnskap om ny teknologi.