

Oracle Best Practices Enterprise Search

Meta
Time: 23 Aug 2006 10:45AM - 11:30AM CST
People: Snehanshu Shah ~24 attendees

Unstructured data
DB artifacts: blobs, VARCHAR, spreadsheets
inside company: can't find content to re-use, code, reports, security, insurance, claims, warranty, IBM anecdote

Behind-the-firewall search
no links, unlike public web, little meta-data, content management, document indexing, intelligent search, text analysis/data mining, case study: Enron emails online and available

Oracle Text
formats: ASCII, HTML, XML, Word, PDF, email, various servers, 150+ formats
Document Services: token identification, themes, gists, classification, clustering

Tokens
Themes: Finds keywords, find tokens, Oracle library has 1M theme mappings, customer: Toro, Themes and Gists: gives context

Questions

Performance: Fastest 'TREK' benchmarks, 200 Gigs took 8 minutes to index on his laptop
Challenges?: pulling out a PDF BLOB is always slow, Speed problem is retrieval once you've found it
Accuracy?: based on tokens, 100%, based on themes
Storage?: Can stay in file systems, use index, even thought it can get stale
Formats?: XML DB and text leverage each other, XML
SES with multiple identities?: You can set up 'impersonations' aliases for users

Secure Enterprise Search (SES)

Issues with Enterprise Search: Finding managers salary, The Numbers hidden in plain view, Manage access by roles and policy
What is SES: Full stack, Install, Data Sources, Logging of searches, IdM integrates with rest of Oracle, 150 Unicode languages, 24/7 support
Customers: Find CRM, Call center logs, transactions, Customer lookup, Fidelity, Subtopic: AT Kerney
UI: Web based, You have to log into the search, Looks like: Weird mix of Google and Oracle Style, Cross link into apps

Application Development

API is simple, wizards available, info on OTN, while it's a big API

Usage

create index, 'select id from table where contains (column, 'cat')>0, discards 'filler words' like 'the' 'on', uses word position lookup, NEAR: Wal-mart.com, K-Mart bluelight online, ArsDigita, Der Spiegel, FlipDog.com, NewsEdge, 170 Systems, SingingFish.com, a few others on slide

Visualization

StretchViewer, Topology Viewer (iview), ThemeMap

