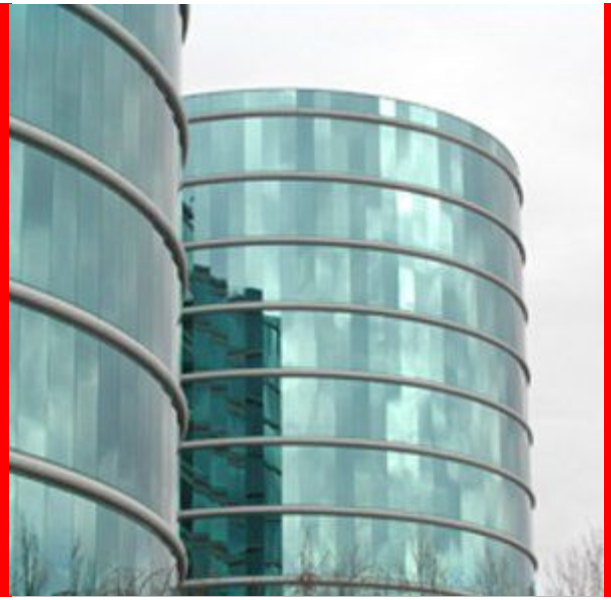
The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions.
The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.



ORACLE®

Oracle Text 11g

Arne Brüning
Leitender Systemberater
arne.bruening@oracle.com

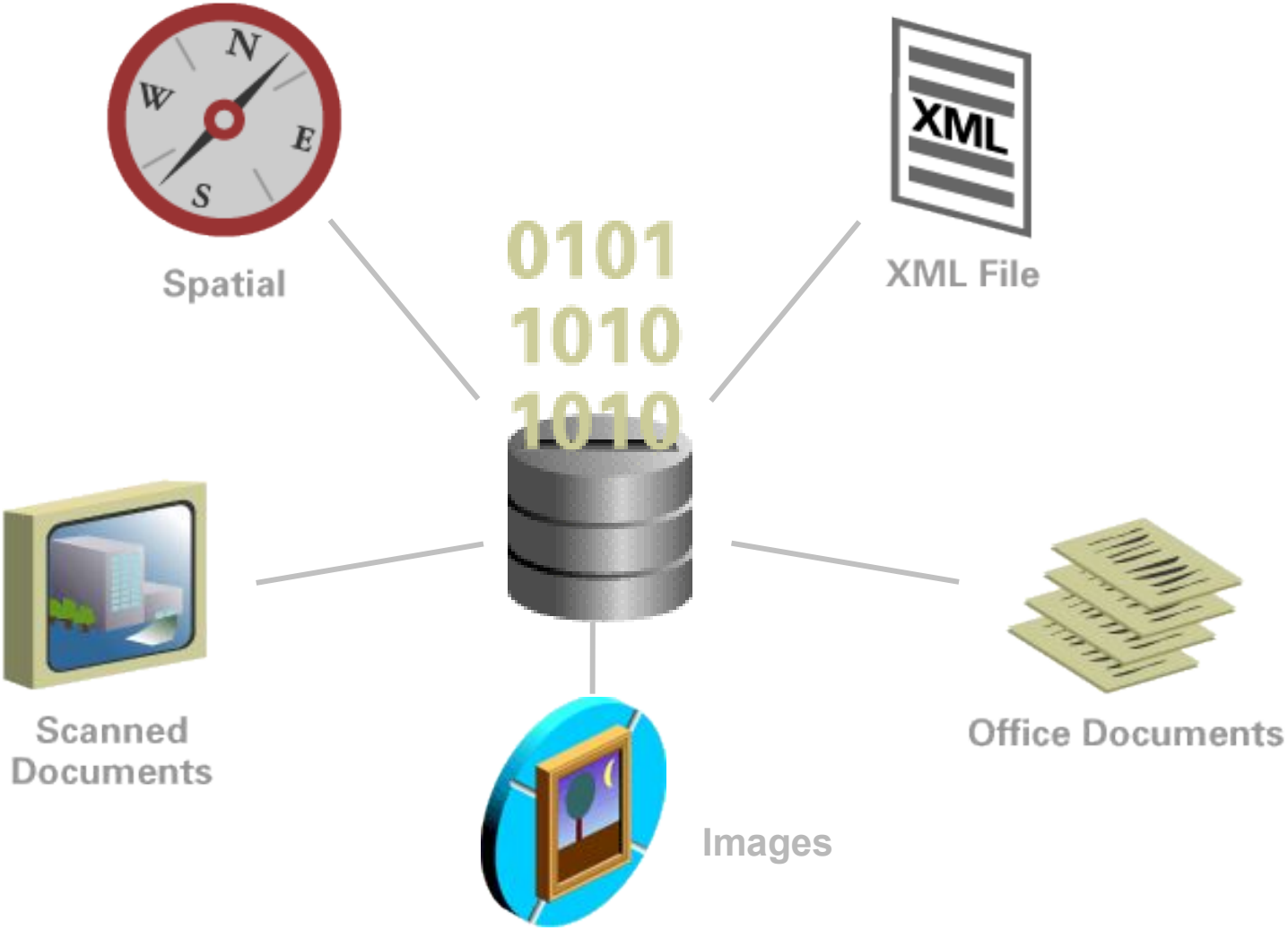
Agenda Oracle Text

- Was ist Oracle Text?
  - … und was ist es nicht?
  - Grundlagen
- Spezielle Features
  - Thesaurus
  - Classification
  - Clustering
- Neue Features in 11g

Oracle's Business

- Oracle Database
  - Manages all kind of data
- Oracle Fusion Middleware
  - Technology Infrastructure for SOA Applications
- Oracle Applications
  - Protect, Extend, Evolve through Fusion Architecture

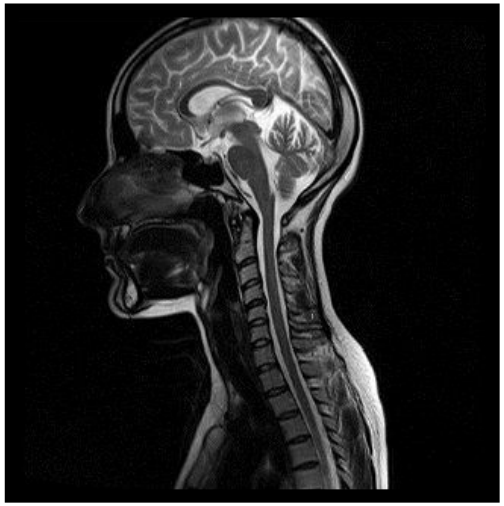
Integrating Unstructured Data



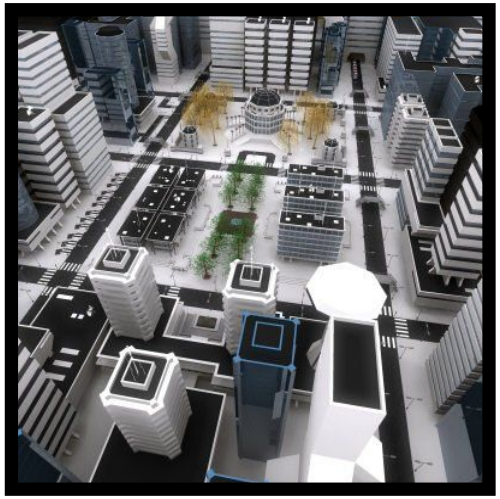
New in Oracle Database 11g

Critical New Data Types

RFID
Data Types



DICOM
Medical Images



3D Spatial
Images

What is Oracle Text?

- “The best kept secret in Oracle”
- Oracle’s information retrieval platform
- Built into the Oracle Database
- Technologies include
  - Free Text Search
  - Natural Language Processing
  - Clustering and Classification
- Oracle Text is included free in EE, SE, and XE

What is Oracle Text - continued

- Oracle Text can index text
  - In the database: VARCHAR2, CLOB, BLOB
  - In the file system (file names held in the database)
  - On the web (URLs held in the database)
  - In many languages
- Text can be
  - Short strings (product names, descriptions)
  - Full sized documents (web pages, emails)
  - Plain text, HTML or proprietary formats (.doc, .pdf)
- Text indexes
  - Are created using CREATE INDEX…
  - Are searched using the CONTAINS clause in SQL
  - Are stored in secondary objects (tables) within the database

Oracle Text Features overview

- *All classical full-text search features...*
  - Exact word matching; Booleans; Wild-cards; 'Fuzzy' matching; Proximity searching ; Stemming in multiple languages ; ISO Thesaurus ; support for Japanese, Chinese, Korean, Western languages
- *Plus Advanced Capabilities...*
  - Linguistic processing to generate themes and gists from text using one million word knowledge base.
  - Advanced ABOUT search
  - Clustering and Classification Features
    - Sorts documents into pre-defined categories
    - Groups documents with similar content
  - Advanced XML search

Extensibility

- Flexible plug-in architecture
- Users can customize
  - Datastore – where the data comes from
  - Filters – how formated documents are translated to indexable text
  - Lexer – how text is broken into words, and how stems or variations of those words are indexed

Oracle Text – A simple example

```
create table simple (id number, text varchar2(2000));

insert into simple values (1, 'the quick brown fox');

create index simple_text on simple (text)
  indextype is ctxsys.context parameters ('');

select text, score(1) from simple
  where contains (text, 'fox') > 0
  and id < 2;

TEXT                                                SCORE(1)
--------------------------------------------------- ----------
the quick brown fox                                          3
```

Oracle Secure Enterprise Search

File Edit View History Bookmarks Window Help

http://abruenin-de2.de.oracle.com:5555/search/query/search

Google

Help

ORACLE®

All Web File Mail

secure enterprise search

Advanced Search
Browse

Search

Help

Powered by Oracle Secure Enterprise Search.
Copyright © 2006, 2007, Oracle. All rights reserved.

Oracle Secure Enterprise Search - secure enterprise search - Mozilla Firefox 3 Beta 1

File Edit View History Bookmarks Tools Help

http://abruenin-de2.de.oracle.com:5555/search/query/search?search.timezone=-60&

Wikipedia (Deutsch)

Hilfe

ORACLE

Alle Web File Mail

secure enterprise search Suchen Erweiterte Suche Durchsuchen

Ergebnisse 1 - 10 von etwa 559 Übereinstimmungen für **secure enterprise search**.

Gruppieren nach: (keine) Sortieren nach: Relevanz

OSES Deutsch: Oracle **Secure Enterprise Search** 10.1.8.1 verfügbar
Dieser Blog beschäftigt sich mit Oracle **Secure Enterprise Search**. Es gibt Informationen zu neuen Versionen
Quellgruppe: **Web** Pfad: oses-d.blogspot.com/2007/05
oses-d.blogspot.com/2007/05/oracle-secure-enterprise-search-10181.html - 77 KB - 22.10.2007 - Gecacht Links

...Ähnliche Dokumente

OSES Deutsch: Mai 2007
Dieser Blog beschäftigt sich mit Oracle **Secure Enterprise Search**. Es gibt Informationen zu neuen Versionen
Quellgruppe: **Web** Pfad: oses-d.blogspot.com
oses-d.blogspot.com/2007_05_01_archive.html - 86 KB - 22.10.2007 - Gecacht Links

...Ähnliche Dokumente

OSES Deutsch
Dieser Blog beschäftigt sich mit Oracle **Secure Enterprise Search**. Es gibt Informationen zu neuen Versionen
Quellgruppe: **Web** Pfad: oses-d.blogspot.com
oses-d.blogspot.com/ - 91 KB - 22.10.2007 - Gecacht Links

...Ähnliche Dokumente

OSES Deutsch: Oktober 2007
Dieser Blog beschäftigt sich mit Oracle **Secure Enterprise Search**. Es gibt Informationen zu neuen Versionen
Quellgruppe: **Web** Pfad: oses-d.blogspot.com
oses-d.blogspot.com/2007_10_01_archive.html - 79 KB - 22.10.2007 - Gecacht Links

...Ähnliche Dokumente

Done

Agenda Oracle Text

- Was ist Oracle Text?
  - … und was ist es nicht?
  - Grundlagen
    - Index Erstellung
    - Abfragen
    - Index-Pflege
- Spezielle Features
  - Thesaurus
  - Classification
  - Clustering
- Neue Features in 11g

The Inverted Index

ID	DOCUMENT
1	the cat sat on the mat
2	the cat in the hat
3	the hat mat

document table

tokenize

1	2	3	4	5	6
the	cat	sat	on	the	mat

discard stop words

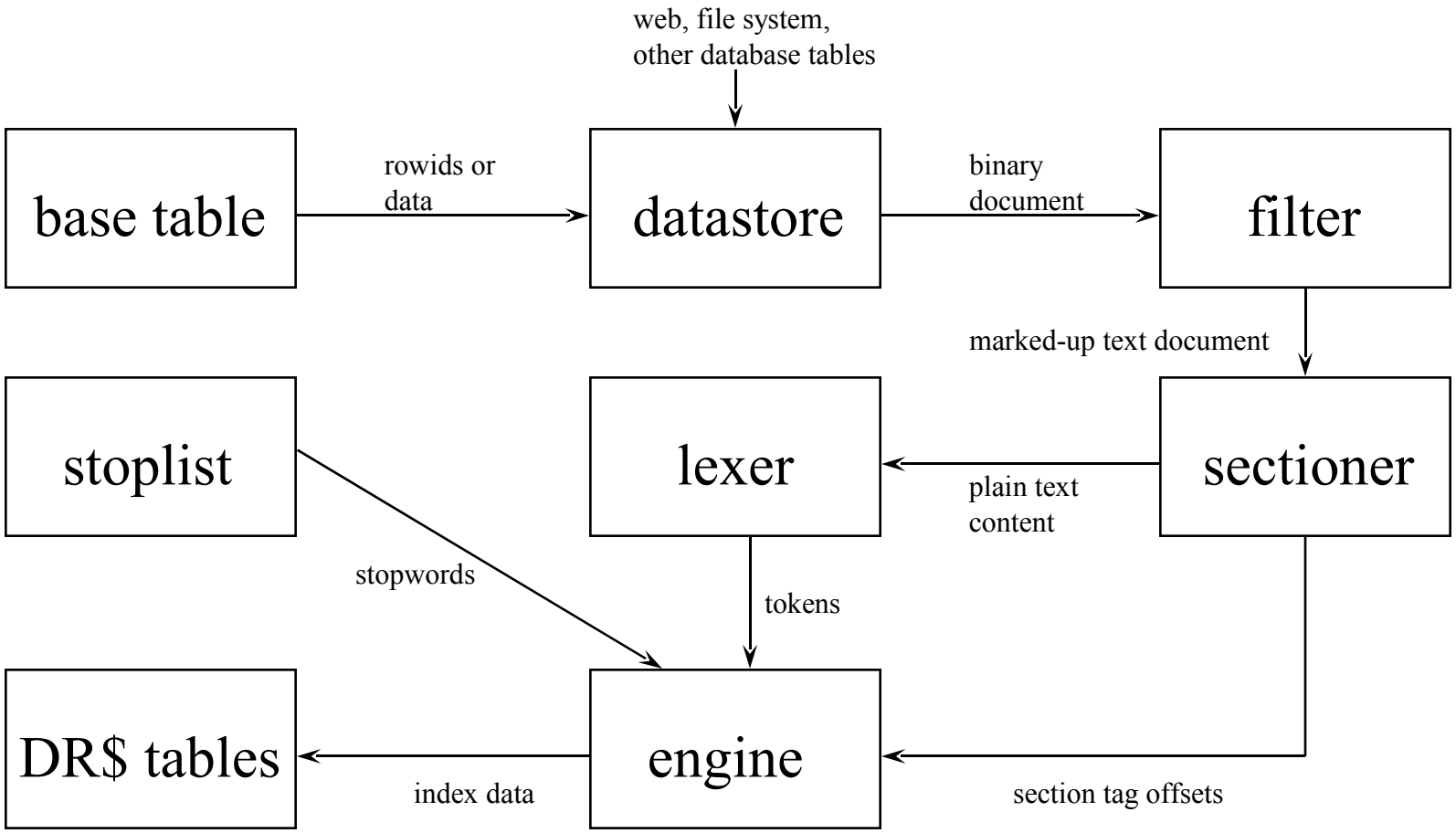
CAT	WORD 2
SAT	WORD 3
MAT	WORD 6

aggregate across documents

TOKEN	DATA
CAT	DOC 1 POS 2 DOC 2 POS 2
SAT	DOC 1 POS 3
MAT	DOC 1 POS 6 DOC 3 POS 3
HAT	DOC 2 POS 5 DOC 3 POS 3

inverted index

The Indexing Pipeline



Indexing Objects: Datastore

- DETAIL_DATASTORE
  - documents are stored in a detail table
  - preference attributes control how to find matching detail table rows
- NESTED_DATASTORE
  - documents are stored in a nested table column
- MULTI_COLUMN_DATASTORE
  - multiple columns of the table are concatenated together
  - w/section searching, allows search across multiple columns with one index

USER_DATASTORE

```
conn ctxsys/ctxsys
create or replace procedure doarc(
  r in rowid,
  c in out nocopy clob
) is
  l_src varchar2(3);
  l_id  number;
  l_con varchar2(2000);
begin
  select src, id into l_src, l_id from AllDoc where rowid = r;
  if (l_src = 'US') then
    select con into l_con from USDoc where id = l_id;
  else
    select con into l_con from UKDoc where id = l_id;
  end if;
  dbms_lob.writeappend(c, length(l_con), l_con);
end;
/
grant execeute on doarc to textuser;
```

Index Objects: Stoplist

- Stoplist is a list of words which do not need to be indexed
- Uses a special API:

```
ctx_ddl.create_stoplist('mysl','BASIC_STOPLIST');
ctx_ddl.add_stopword('mysl','the');
```

- BASIC_STOPLIST
  - list of words for mono-lingual corpora
- MULTI_STOPLIST (9.0.1)
  - list of language-specific stopwords
- Stoplist Enhancements (8.1.6)
  - Support for Stopthemes and Stopclasses in Stoplists
  - Dynamic Addition of Objects To Stoplists

Index Objects: Lexer

- MULTI_LEXER (8.1.6)
  - supports heterogenous languages
- USER_LEXER (9.2)
  - user-supplied PL/SQL procedures to tokenize and normalize
- WORLD_LEXER (10g)
  - UNICODE-based lexer that follows different strategies for different languages based on autorecognition by codepoint range

Multi-Lingual Corpora

- WORLD_LEXER (10g)
  - UNICODE-based lexer which varies tokenization strategy by codepoint analysis
    - whitespace segmentation for European languages, VGRAM for Asian languages, does some basic segmentation for Arabic, etc.
    - Easier to set up than MULTI_LEXER
    - Currently no attributes, so you get what you get
    - Area of future development
- UTF-16 Auto-detection (Little / Big Endian) (9.0.1)

Index Objects: Filter

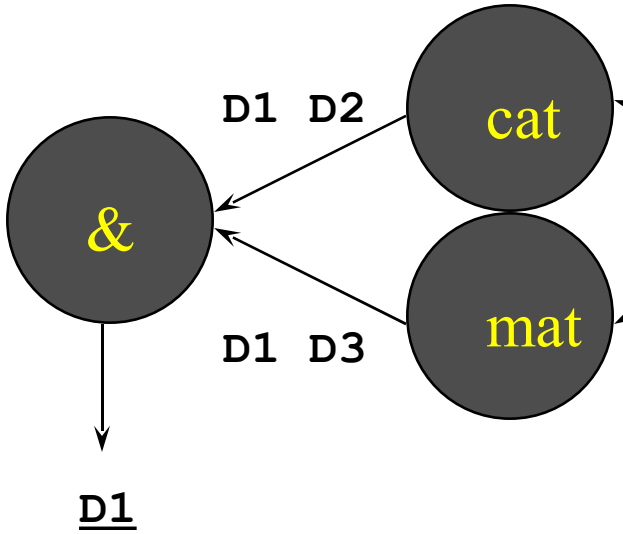
- INSO_FILTER
  - Filters 100+ binary formats including PDF and MS Office to text
  - Relies on an executable “ctxhx” which uses third-party code from Stellent
  - Resource-intensive
- In 10gR2 (and 9.2.0.7+, 10.1.0.4+)
  - AUTO_FILTER: New filter vendor, faster, more formats
- PROCEDURE_FILTER
  - User-supplied PL/SQL procedure to filter

Agenda Oracle Text

- Was ist Oracle Text?
  - … und was ist es nicht?
  - Grundlagen
    - Index Erstellung
    - Abfragen
    - Index-Pflege
- Spezielle Features
  - Thesaurus
  - Classification
  - Clustering
- Neue Features in 11*g*

Querying an Inverted Index

query: CAT AND MAT



TOKEN	DATA
CAT	DOC 1 POS 2 DOC 2 POS 2
SAT	DOC 1 POS 3
MAT	DOC 1 POS 6 DOC 3 POS 3
HAT	DOC 2 POS 5 DOC 3 POS 3

inverted index

Querying the Index

- Query using the CONTAINS clause:

```
select * from foo
    where contains(text, 'queryterm')>0
```

  - first argument is column name, second argument is query term
  - use anywhere select can be used
  - supports all database generic query features

Querying the Index

- Relevance ranking
  - SCORE operator returns a number characterizing relevance of the document to the query
  - Link SCORE to the CONTAINS using ancillary data label:

```
select score(1), id from foo
where contains(text, 'queryterm',1)>0
order by score(1) desc
```

  - Score algorithm is a variant of TF/IDF, affected by popularity of term in document and in corpus, and number of documents in the index.

Context Query Language

- Term/keyword
  - looks for documents containing this word
  - wise to surround your term in curly braces to avoid conflict with operators and reserved words:

```
contains(text, '{someword}')>0
```

- Phrase
  - no special delimiters needed to signify a phrase

Context Query Language

- Expansion Operators
  - wildcard (%, _), fuzzy (?), stem ($), soundex (!)
  - work by expanding the pattern and transforming the query into essentially a big OR
  - large expansions may slow because of 1000's of terms

Context Query Language

- Proximity
  - `dog ; cat`
  - `NEAR((dog, cat, pig), 10)`
- ABOUT (engl.)
  - with theme indexing, does thematic search
  - `about(railroads)`
- Thesaurus operators
  - SYN, BT, NT, etc.
  - `SYN(dog, mythesaurus)`
  - user must provide and load the thesaurus -- not built-in

Orthography: Diacritics

- Changes in form due to diacritics (schwül, schwul)
- Generally a cross-language search problem
  - Diacritic marks are not disposable within a language
  - Non-native speakers may drop the diacritics in query
  - Should allow such query to find word in corpus
- BASIC_LEXER includes the BASE_LETTER attribute
  - when set, will normalize characters with diacritics to base forms without diacritics

Orthography: Alternate Spelling (8.1.5)

- Standardized variant spelling for foreigners
  - example: Tüte > Tuete, oppebær > oppebaer
- compound characters
  - example: ißt > isst
- BASIC_LEXER ALTERNATE_SPELLING implements normalization for a specific language's set of variant orthography
  - choices: GERMAN, DANISH, SWEDISH
  - will index words twice: once with ß, once with ss e.g.

Inflection

- Inflection
  - noun plurals
  - Some languages have declension of nouns
- Inflection is handled through the stem operator
  - example: contains(a, ‘$apple’)>0 finds apple, apples
  - done through expansion
  - lexical software from InXight
  - stemmer is set in the wordlist at create index time, but only really has effect at query time

Decompounding

- Some whitespace-delimited languages have widespread compound terms
  - German is the main culprit: Rechtschreibreform, Nordhauptbahnhof, etc.
  - Search for “bahnhof” should hit Nordhauptbahnhof
- BASIC_LEXER attribute COMPOSITE, can be set to GERMAN or DUTCH
  - each word passed through decompounder
    - splits the token into multiple tokens, possibly overlapping
    - Nordhauptbahnhof-> nord, haupt, bahnhof, hauptbahnhof

Segmentation

- Japanese and Chinese do not use whitespace
- Two strategies:
  - VGRAM: split text into overlapping segments
    - ABCD > AB, BC, CD e.g.
    - query for “ABC” queries for the phrase “AB BC”
    - always works, but it slow and produces tons of tokens
  - Lexicon: use a dictionary and greedy match
    - ABCD > ABC D, if ABC is a word
    - query for “ABC” looks for “ABC”
    - produces fewer tokens, works like western IR, but not 100%

Agenda Oracle Text

- Was ist Oracle Text?
  - … und was ist es nicht?
  - Grundlagen
    - Index Erstellung
    - Abfragen
    - Index-Pflege
- Spezielle Features
  - Thesaurus
  - Classification
  - Clustering
- Neue Features in 11*g*

Maintaining the Index: DML

- Context indexes are not transactional
  - structure is inherently aggregate
  - difficult and expensive to update
- Inserts and updates are delayed addition to index
- Documents waiting to be indexed are stored in queue
- Synchronization adds new and updated documents to the index
  - memory Parameter (9.0.1)

```
ctx_ddl.sync_index('indexname');
```

Maintaining the Index: Optimize

- incremental update in sync fragments the index
- what is fragmentation?

after create idx	`CAT D1 D2`	
sync	`CAT D1 D2` `CAT D3`	this is sub-optimal, so
optimize	`CAT D1 D2 D3`	

Maintaining the Index: Optimize

- why optimize?
  - makes query faster
    - fewer rows = less I/O
    - data is more efficiently stored = smaller data = less I/O
    - data is more localized
  - recover wasted space
    - deleted and updated documents are not removed from the index
    - optimize lazy-deletes the data from the index

Maintaining the Index: Optimize

- Recommend: Full optimize

```
ctx_ddl.optimize_index('myindex','FULL',maxtime=>10)
```

- optimizes as many rows as possible in 10 minutes
- if time runs out, saves state so next invocation can pick up where it left off
- optimize is rewriting rows, so can take up more time and REDO/UNDO than index creation
- for large systems, can be done in parallel

Maintaining the Index: Optimize

```
ALTER INDEX textidx rebuild;

ALTER INDEX newsindex rebuild parameters('replace
lexer my_lexer');
```

- REBUILD optimize (10g)
  - rewrites the entire index table using direct path load
  - can complete optimization on entire index faster than FULL method, with less REDO/UNDO

Maintaining the Index: DML

- suggest setting up a dbms_job to call sync periodically
  - how frequently? as rarely as is feasible
- SYNC AUTOMATIC at create index sets up a sync job for you (10g)
- SYNC ON COMMIT does an automatic sync after each commit (10g)
  - this may greatly increase fragmentation
  - consider TRANSACTIONAL

Maintaining the Index: DML

- TRANSACTIONAL (10g) enables transactional query semantics
  - records unindexed rowids
  - query joins a function scan on unindexed rowids with index results
  - will be slower than normal query
  - can be turned off in a session; consider using non-transactional for queries that don't need transactional semantics

Agenda Oracle Text

- Was ist Oracle Text?
  - … und was ist es nicht?
  - Grundlagen
- Spezielle Features
  - Thesaurus
  - Classification
  - Clustering
- Neue Features in 11g
- Appendix

Document Services

- Filter a binary document to text
- Highlight text query hit words in a document
- Document summarization by key sentence/paragraph extraction
- Main themes extraction of a document (from built-in knowledge-base)
- Keyword in Context (KWIC) (10.2)
- Package name: `ctx_doc`

Index Objects: Section Group

- XML_SECTION_GROUP (8.1.6)
  - XML tagging
  - This is Not an XML parser. Does not validate or support advanced XML features
  - add sections dynamically after indexing with ALTER INDEX
  - **XMLType Indexing**
- AUTO_SECTION_GROUP (8.1.6)
  - like the XML_SECTION_GROUP, but automatically indexes every tag as a ZONE section
  - add sections dynamically after indexing with ALTER INDEX.
- PATH (9i)
  - like ZONE, but supports XPath-like queries
- PATH_SECTION_GROUP (10g)
  - like the AUTO_SECTION_GROUP, but indexes every tag as a PATH section

Context Query Language

- WITHIN (8.1.5, hierarchical 8.1.6)
  - limits search to a particular zone or field section of the section group
- HASPATH / INPATH (9.0.1)
  - does simple Xpath-like searches
    `dog INPATH(/A/B//C[/D = "animal"])`
  - Highlightning (10g)
- MDATA (10g)
  - Searches for MDATA section values
    `MDATA(author, william shakespeare)`

Query Template (9.2)

Main idea:

- XML-like language for complex queries:

```
contains(text, '
<query>
  <textquery>cat or dog</textquery>
  <score datatype="float"/>
</query>
')>0
```

  - override grammar, control score, query language etc.
- Progressive Relaxation (10.2)

Ohne Progressive Relaxation

```
select * from pr where contains( doc, 'Arne Brüning')>0;

select * from pr where contains( doc, 'near((Arne,
Brüning), 1)')>0;

select * from pr where contains( doc, 'Arne and
Brüning')>0;
```

Progressive Relaxation

```
select * from pr where CONTAINS (doc,
  '<query>
    <textquery lang="GERMAN" grammar="CONTEXT">
      <progression>
        <seq>{Arne} {Brüning}</seq>
        <seq>{Arne} NEAR {Brüning}</seq>
        <seq>{Arne} AND {Brüning}</seq>
      </progression>
    </textquery>
    <score datatype="INTEGER" algorithm="COUNT"/>
  </query>'
)>0;
```

ISO-Konformer Thesaurus

Thesaurus Manager fuer Oracle: Webanwendung by moving objects - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Zurück Suchen Favoriten

THESAURUS MANAGER FÜR ORACLE

© 2005 moving objects

Thesaurus: BE Begriff: kaumann%

- Akademiker
  - Engere Begriffe
    - Akademikerin
    - Akademikerverband
    - Altertumsforscher
    - Architekt
    - Betriebswirt
      - Weitere Begriffe
        - Akademiker
        - Wirtschaftswissenschaftle
      - Engere Begriffe
        - Agrarbetriebswirt
        - Holzbetriebswirt
        - Sportökonom
        - Technischer Betriebswirt
        - Verkehrsbetriebswirt
      - Synonyme
        - Betriebswirtschaftler
      - Zugehörige Begriffe
        - Volkswirt
          - Weitere Begriffe
          - Synonyme
            - Statistiker
              - Weitere Begriff
              - Engere Begriffe
                - Demoskop
                - Weitere

Begriff bearbeiten

Thesaurus: BE

Begriff: Betriebswirt

Attribute | Beziehungen | Hinweis

Homonym-Zusatz:

Bevorzugter Begriff:

Bevorzugter Begriff?

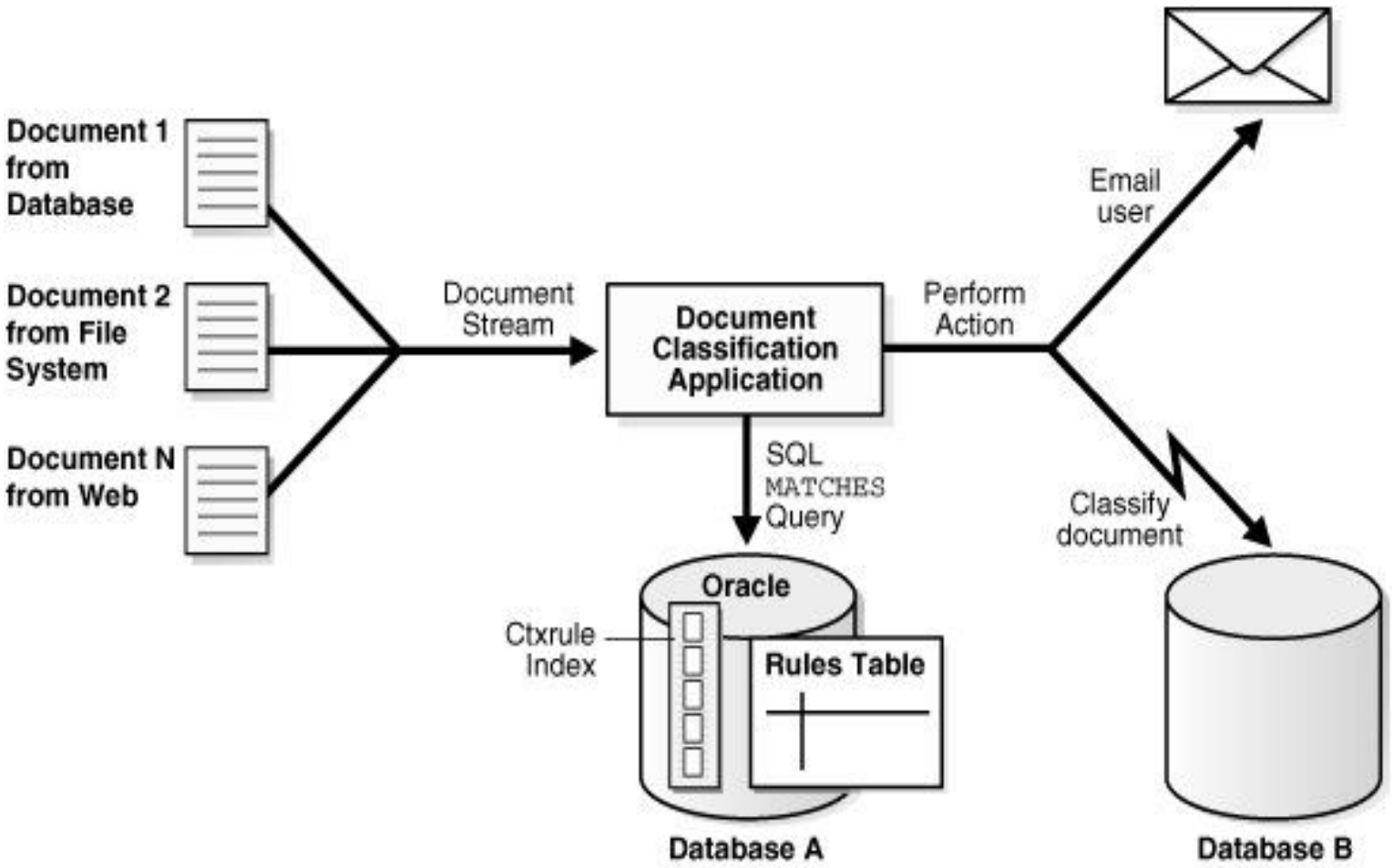
Geltungsbereich-Hinweis:

ThesMan Term Attribu

Speichern Abbrechen

Lokales Intranet

Classification



Classification in Oracle Text

- Example with **`ctxrule`**

```
insert into qry values (1,'cat & mat');
insert into qry values (2,'cat & dog');

create index qryx on qry(q)
    indextype is ctxsys.ctxrule;

select id from qry
    where matches(q, 'the cat sat on the mat')>0
    returns "1"
```

Classification in Oracle Text

- Classification is the next step up from routing
  - given a corpus of documents organized into related groups, create rules to route new documents to correct groups (9i)
  - **`ctx_cls.train`** (9.2)
  - output is a list of queries which can be fed into **`ctxrule`**
  - use decision tree or support vector machines (10g)

Oracle Text bei Gruner & Jahr

Verschlagwortung - Mozilla Firefox

Datei Bearbeiten Ansicht Gehe Lesezeichen Extras Hilfe

Verschlagwortung

Gruner+Jahr AG & Co KG
Druck- und Verlagshaus
Hamburg

Lektoratsclient Version 1.1

www.guj.de

FRANKFURTER ALLGEMEINE ZEITUNG WIRTSCHAFT 01.11.2005 SEITE: 9

ED#: D9W781P0
597 WÖRTER
KEINE GRAFIKEN

STATUS: NICHT FREIGEGEBEN

Infineon und IG Metall einigen sich auf Schließung des Münchner Werks

Nach acht Tagen Streik steht ein Sozialtarifvertrag fest / 615 Mitarbeiter verlieren 2007 ihren Arbeitsplatz

him

Bibliographische Informationen

Verschlagwortung

Themenbereiche 'Arbeit und Sozialstaat' 'Industrie'

Klassifikationen

Länder

Sachthemen 'Gewerkschaft' 'Streik' 'Metallindustrie'

Personen

☐ Keine männliche oder weibliche Person ☐ Männliche Person ☐ Weibliche Person

Organisationen

Automatische Klassifizierung

- ☑ **Arbeit und Sozialstaat (93)**
  - ☑ **Gewerkschaft (57)**
  - ☑ **Streik (54)**
  - ☐ Tarifpolitik (36)
- ☑ **Industrie (84)**
  - ☑ **Metallindustrie (58)**
  - ☐ Elektronisches Bauteil (49)

☐ Otto Wiesheu (2)
☐ Werner Neugebauer (2)

☐ Infineon Technologies AG (12)
☐ IG Metall (6)
☐ CSU (2)

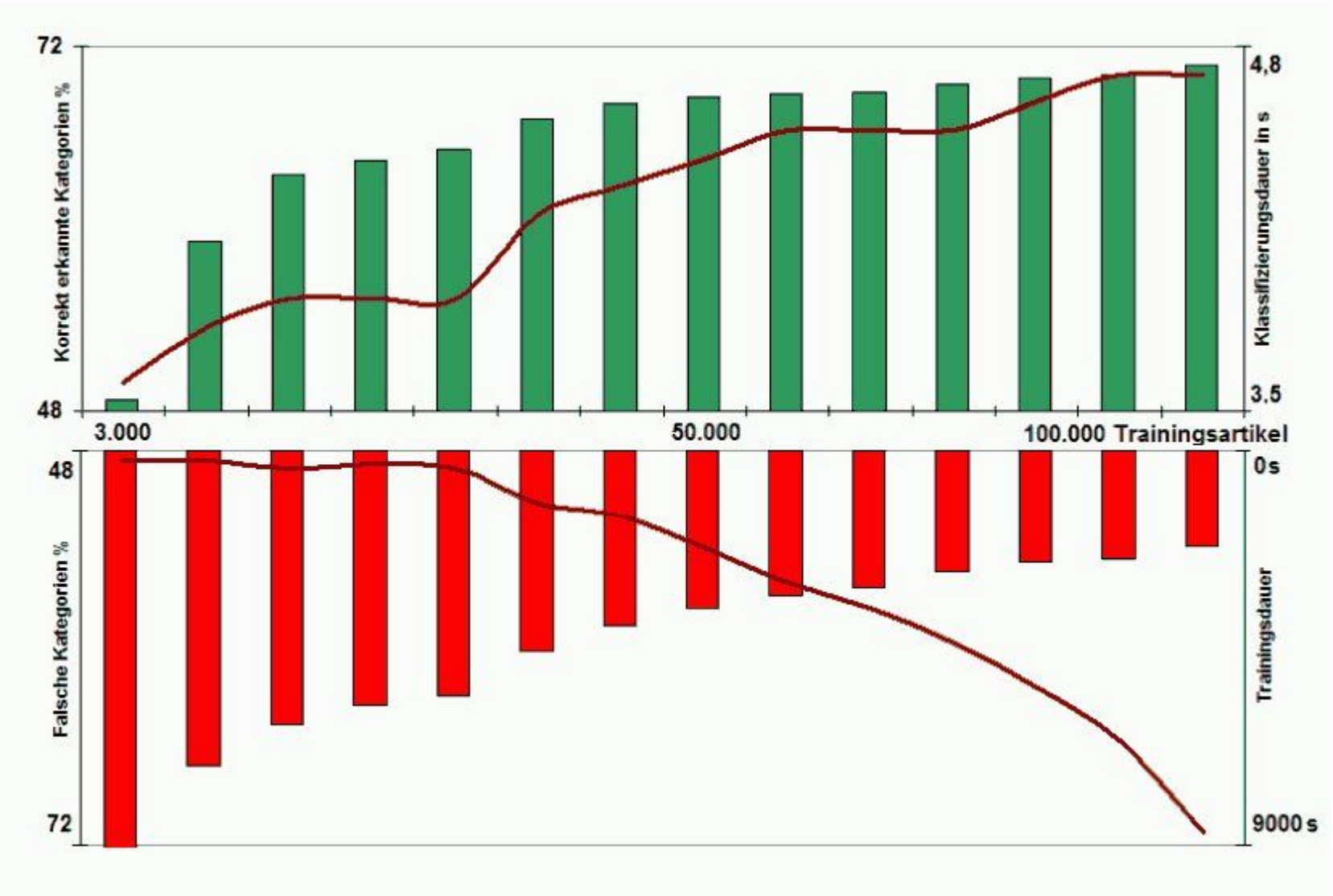
Volltext | Personen | Organisationen | Templates | DigDoc | << | >> | Speichern | Abbrechen

ORACLE®

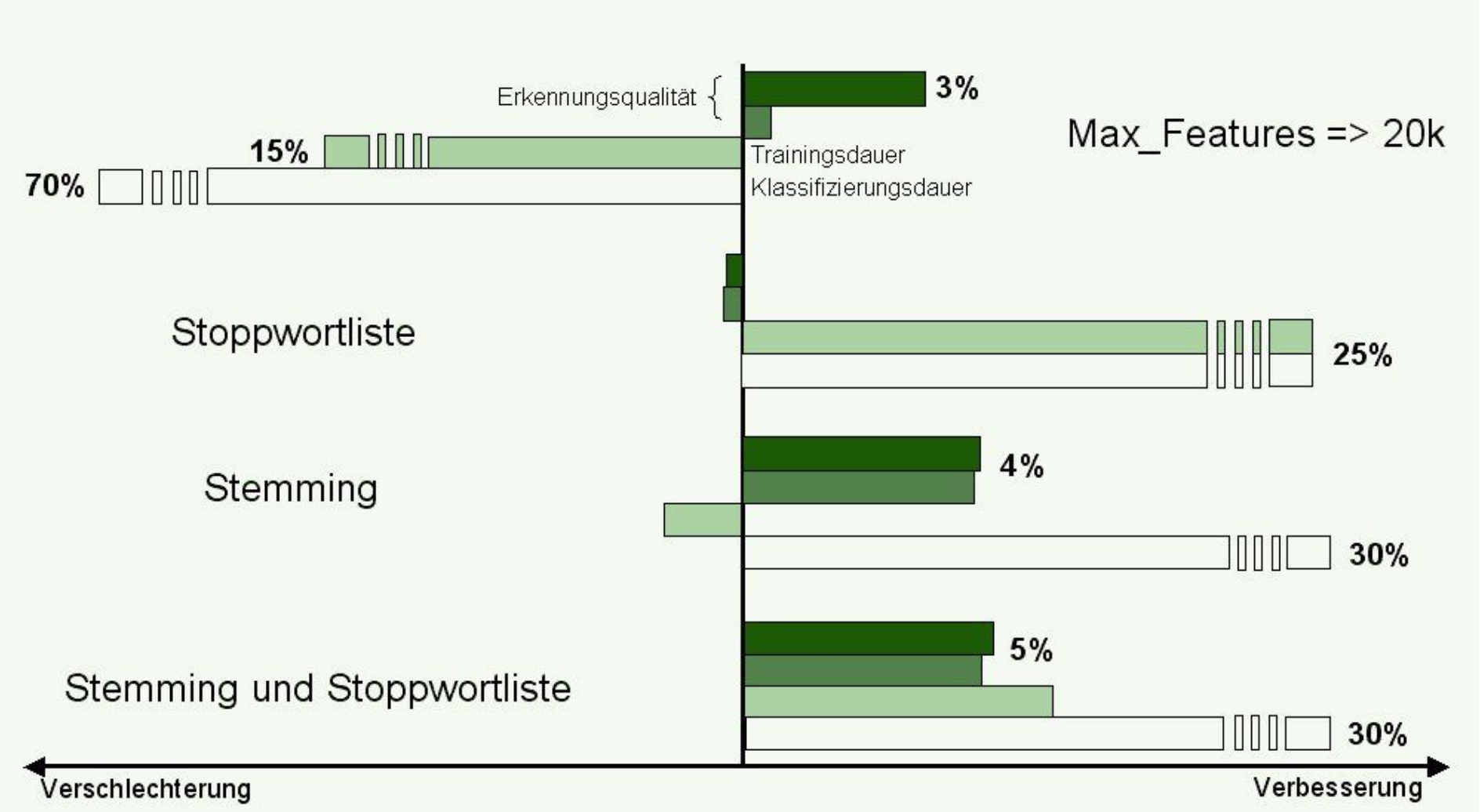
D E M O N S T R A T I O N

Classification

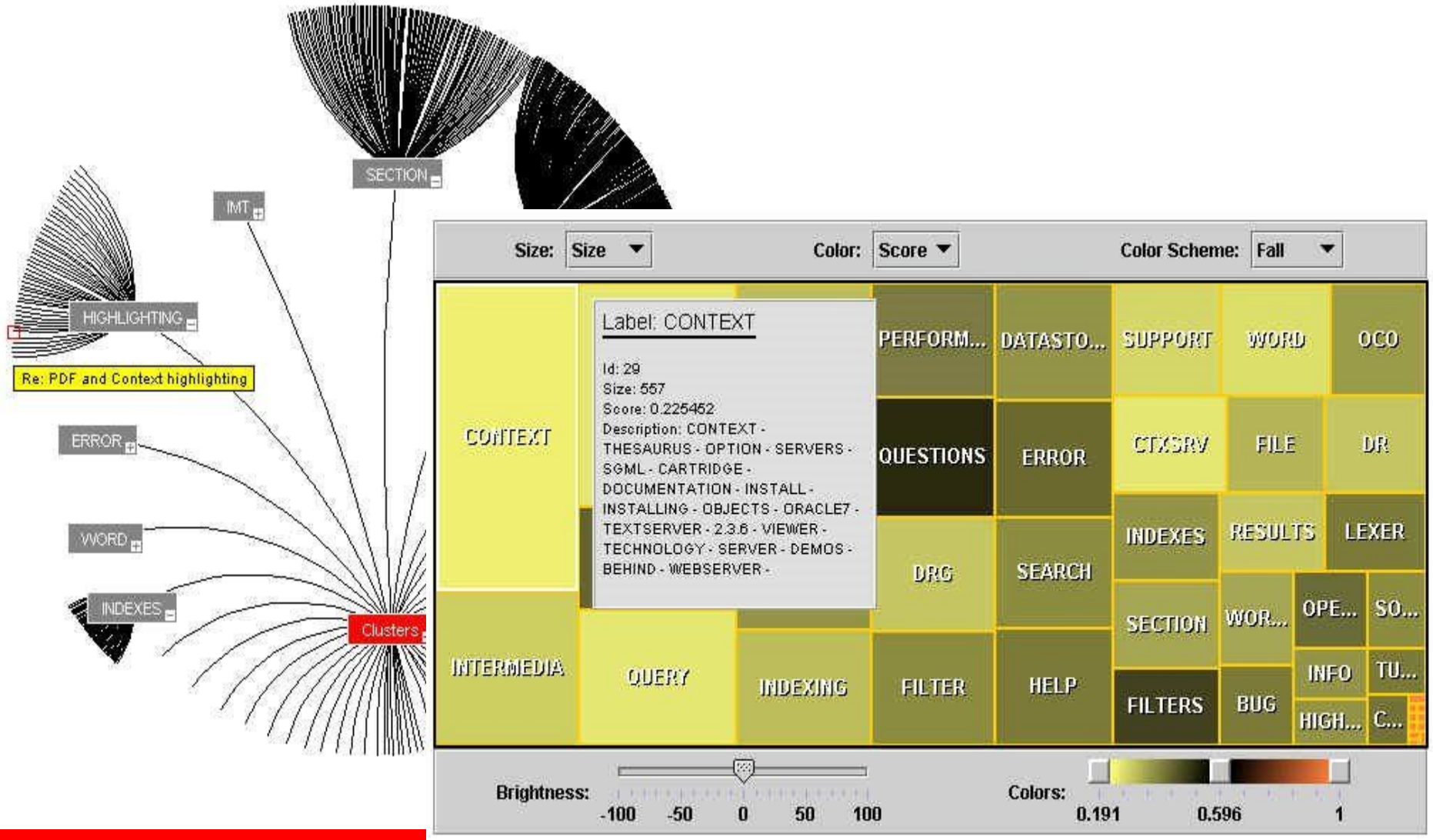
Anzahl Trainingsartikel



Parameterwahl



Clustering



Oracle Secure Enterprise Search - secure enterprise search - Mozilla Firefox 3 Beta 1

File Edit View History Bookmarks Tools Help

http://abruenin-de2.de.oracle.com:5555/search/query/search?search.timezone=-60&

Wikipedia (Deutsch)

Hilfe

ORACLE

Alle Web File Mail

secure enterprise search Suchen Erweiterte Suche Durchsuchen

Ergebnisse 1 - 10 von etwa 559 Übereinstimmungen für **secure enterprise search**.

Gruppieren nach: (keine) Sortieren nach: Relevanz

Ergebnisse filtern nach Ausblenden

- Thema (100)
  - oracle (41)
    - oracle database (6)
    - oracle corporation (5)
    - oracle fusion middleware (3)
    - diverse (28)
  - enterprise manager (10)
    - enterprise manager related (3)
    - diverse (7)
  - oracle webcenter (9)
    - webcenter suite (3)
    - diverse (6)
  - management (25)
  - oracle secure enterprise search (32)
    - blog beschäftigt sich mit oracle secure enterprise (12)
    - diverse (20)
  - database (21)
  - oses deutsch (12)
  - search oracle (8)
  - secure backup (8)
  - sample code tutorials oracle

OSES Deutsch: Oracle **Secure Enterprise Search** 10.1.8.1 verfügbar

Dieser Blog beschäftigt sich mit Oracle **Secure Enterprise Search**. Es gibt Informationen zu neuen Versionen

Quellgruppe: Web Pfad: oses-d.blogspot.com/2007/05

oses-d.blogspot.com/2007/05/oracle-secure-enterprise-search-10181.html - 77 KB - 22.10.2007 - Gecacht Links

...Ähnliche Dokumente

OSES Deutsch: Mai 2007

Dieser Blog beschäftigt sich mit Oracle **Secure Enterprise Search**. Es gibt Informationen zu neuen Versionen

Quellgruppe: Web Pfad: oses-d.blogspot.com

oses-d.blogspot.com/2007_05_01_archive.html - 86 KB - 22.10.2007 - Gecacht Links

...Ähnliche Dokumente

OSES Deutsch

Dieser Blog beschäftigt sich mit Oracle **Secure Enterprise Search**. Es gibt Informationen zu neuen Versionen

Quellgruppe: Web Pfad: oses-d.blogspot.com

oses-d.blogspot.com/ - 91 KB - 22.10.2007 - Gecacht Links

...Ähnliche Dokumente

OSES Deutsch: Oktober 2007

Dieser Blog beschäftigt sich mit Oracle **Secure Enterprise Search**. Es gibt Informationen zu neuen Versionen

Done

ORACLE®

D E M O N S T R A T I O N

Clustering

Recap of classification and clustering

- Classification
  - Supervised classification of content
  - Two ways: rules or training sets
  - You can group a number of categories into a taxonomy
  - Very useful for defining a common vocabulary in an enterprise
- Clustering
  - Unsupervised classification of patterns into groups
  - The engine analyzes the document collection and outputs a set of clusters with documents on it
  - Very useful for *discovering* patterns or nuggets in collections
  - Could be used as a starting point when there is no taxonomy present

Agenda Oracle Text

- Was ist Oracle Text?
  - … und was ist es nicht?
  - Grundlagen
- Spezielle Features
  - Thesaurus
  - Classification
  - Clustering
- Neue Features in 11g

Oracle Text 11g

Focus Areas

- Query Performance and Scalability
- Internationalization
- Zero Downtime for Applications

Composite Domain Index

Composite Domain Index – why?

- “Mixed Queries” are a strength and a weakness
- Great flexibility, sometimes not-so-great performance.
- Costly if both text and structured part are non-selective

```
SELECT item_id FROM items
WHERE CONTAINS (description, 'music') > 0
AND type = 'BOOK'
AND price < 10
ORDER BY price
```

Mixed Query Processing

- Look up '*music*' in text index
- Get rowid for each text index hit
- For each row from text index:
  - Check item type (base table lookup or index combine)
  - Check price (base table lookup or index combine)
- Sort results (base table lookup or index scan)

Earlier solutions

- Tagging or Field Sections
  - .. blah blah XXTYPE%book
  - WHERE CONTAINS (description, 'music and xxtype%book')
  - .. blah blah <TYPE>book</TYPE>
  - WHERE CONTAINS (description, 'music and book within itemtype'
- Fast – structured clause satisfied directly from text index
- Does not solve range searching
- Does not solve sort issues
- Change "structured" data -> reindex whole document
- Can be complex to build

MDATA Sections

- M(eta)DATA Sections Introduced in Oracle 10g

```
insert into library_stock values
   (2, '<title>The World According to Garp</title> <author>John Irving</
   author> <status> In Stock</status> <stocklevel>12</stocklevel>');

exec ctx_ddl.add_mdata_section(group_name=>'mysg',
   section_name=>'status', tag=>'status');

select book_info from library_stock where contains (book_info, 'irving
   within author and mdata(status, In Stock)') > 0;
```

- Transactional

  Can update MDATA without reindexing whole document

- Oracle.com => Search for "mdata tips"

MDATA Limitations

- No range searches
- No help with sorting
- So … we could use a new section type for Structured DATA…

Introducing SDATA

```
insert into library_stock values
   (2, '<title>The World According to Garp</title> <author>John
   Irving</author> <status> In Stock</status>
   <stocklevel>12</stocklevel>');

exec ctx_ddl.add_sdata_section(group_name=>'mysg',
   section_name=>'stock', tag=>'stocklevel',
   datatype=>'number');

select book_info from library_stock where contains (book_info,
   'irving within author and sdata(stock > 1)') > 0;
```

Sorting on SDATA

- Relies on new feature: "User Defined Scoring"

```
select book_info from library_stock where contains (book_info,
  '<query>
    <textquery>
      irving within author and sdata(stock > 1)
    </textquery>
   <score normalization_expr = “sdata(stock)”/>
</query>') > 0
```

But …

- What I want…

```
select book_info from library_stock
    where contains (book_info, 'irving') > 0
    and stock > 1
    order by stock
```

- What I have…

```
select book_info from library_stock where contains (book_info,
    '<query><textquery>
       irving within author and sdata(stock > 1)
    </textquery><score normalization_expr =
    “sdata(stock)”/></query>') > 0
```

Composite Domain Indexes solve this

```
CREATE INDEX book_index
  ON library_stock (book_info)
  INDEXTYPE IS CTXSYS.CONTEXT
  FILTER BY stock [, … ]
  ORDER BY stock [, …] [ DESC ];
```

```
select book_info from library_stock
   where contains (book_info, 'irving') > 0
   and stock > 1
   order by stock
```

Composite Domain Index

- “Composite” because the index is composed of multiple columns
- Primary column is free-text indexed. Auxiliary columns are indexed invisibly as SDATA sections
- Query optimizer will "push down" filtering and sorting into the text index when appropriate
- Column types:
  - VARCHAR2(249) (max)
  - RAW(249) (max)
  - Number
  - Date

New Optimizer Hints

```
SELECT /*+ DOMAIN_INDEX_SORT
  DOMAIN_INDEX_FILTER(items items_description) */
   id, description, price
 FROM items
 WHERE contains(description, 'music') > 0
      AND type = 'books'
 ORDER BY price DESC;
```

Benefits

- Avoid DOCID->ROWID translations for intermediate hits which are eliminated from final results
- Fetching of structured info from $S IOT is much faster than fetching from sparse base table blocks

- Some internal benchmark results:
  - Structured predicates: 10x faster
  - Sorting: 4x faster
    - Your results may vary!

Other new Index Features

Recreate Index Online

- Many changes to an index take effect only when documents are reindexed
- Critical applications cannot afford down-time
- Previous solution:
  - Create new user_datastore index on dummy column
  - When complete, change application to point to new index
  - Drop old index
- Works, but cumbersome and error-prone
- Doesn't allow for other datastore types

Recreate Index Online - SQL

- CTX_DDL.CREATE_SHADOW_INDEX
  (idx_name=>'items$description',
  parameter_string=>'REPLACE LEXER
  my_new_lexer');

- CTX_DDL.EXCHANGE_SHADOW_INDEX
  (idx_name => 'items$description'
  [partition_name => 'partname']);

Time-Limited Index Creation

- Creation of an index can be time-limited to avoid slowing down system at peak times

```
CREATE INDEX items$description
  ON items(description)
  INDEXTYPE IS CTXSYS.CONTEXT
  PARAMETERS('NOPOPULATE')

CTX_DDL.POPULATE_PENDING
   (idx_name=>'items_description')

CTX_DDL.SYNC_INDEX
  (idx_name=>'items$description', maxtime=>480);
```

Agenda Oracle Text

- Was ist Oracle Text?
  - … und was ist es nicht?
  - Grundlagen
- Spezielle Features
  - Catalogs
  - Classification
  - Multi-lingua corpora
- Neue Features in 11g

Some Oracle Text Customers

Walmart★com

DER SPIEGEL

NewsEdge

ArsDigita
open for e-business

170 SYSTEMS

welcome 歡迎 欢迎

Oracle Technology Network | Downloads, Discussions, and Documentation for Developers and DBAs

File Edit View History Bookmarks Window Help

http://www.oracle.com/technology/index.html

ORACLE TECHNOLOGY NETWORK

(Sign In/Register for Account | Subscribe)

Oracle Websites

secure search

Technology Network

PRODUCTS
- Database
- Middleware
- Developer Tools
- Enterprise Management
- Applications Technology
- Products A-Z

TECHNOLOGIES
- BI & Data Warehousing
- Embedded
- Java
- Linux
- .NET
- PHP
- Security
- Service-Oriented Architecture
- Windows Server System
- Virtualization
- Technologies A-Z

COMMUNITY
- Join OTN
- Oracle ACEs
- Oracle Wiki
- Blogs
- Podcasts
- Events
- Newsletters
- Oracle Magazine
- Oracle Books
- Certification
- User Groups

shortcuts | GETTING STARTED | DOWNLOADS | DOCUMENTATION | FORUMS | ARTICLES | SAMPLE CODE | TUTORIAL

Printer View | E-mail this page | Bookmark

RECENTLY POSTED

Call for Nominations: 2008 Oracle Excellence Awards

Have you worked on a solution that uses Oracle Applications with Oracle Fusion Middleware in some creative way? Get some recognition via The Oracle Excellence Awards (nomination deadline: Aug. 8).

posted 6/13/08 16:51:22 GMT - Tags: middleware, java, soa

Technical Article: Return to Formsville

It's still important to know the key architectural concepts common to many Oracle Forms-based applications. This Technical Article from Oracle ACE Director Chris Muir and Oracle ACE Penny Cookson explain why.

posted 6/13/08 16:50:53 GMT - Tags: middleware, java, soa

Save the Date: BEA Welcome and Oracle's Middleware Strategy Briefing Webcast, July 1

Join Oracle's Charles Phillips and Thomas Kurian for a briefing about how the addition of BEA products to Oracle Fusion Middleware will create a best-in-class combination. (See also: Welcome, Dev2Dev & Arch2Arch FAQ)

posted 6/2/08 17:21:58 GMT - Tags: middleware, java, soa

▶ MORE HEADLINES ▶ CRITICAL PATCH UPDATES ▶ TAG CLOUD

FEATURED DOWNLOADS
- Oracle Database 11g Release 1
- Oracle JDeveloper 11g Technology Preview 4
- Oracle VM Free product
- Oracle SQL Developer Free product
- Oracle Database XE Free product

▶ MORE DOWNLOADS

DEVELOPER EVENTS

OTN Developer Day - The Fusion Development Experience
7/1/2008 — Redwood Shores, Calif.

OracleDays 2008
7/14/2008 - 7/18/2008 — Bellevue, Wash.

▶ MORE DEVELOPER EVENTS

BLOGS

Come and Get It: Blogger Credential for Oracle OpenWorld
posted 6/13/2008 in OTN TechBlog (Justin Kestelyn)

Oracle's Arch2Arch Newsletter
posted 6/10/2008 in OTN TechBlog (Justin Kestelyn)

Für weitere Informationen

setzen auch wir OSES ein…

http://search.oracle.com

F&A

FRAGEN

ANTWORTEN

ORACLE IS THE INFORMATION COMPANY