TREX – SAP NetWeaver’s Search and Classification Engine

SAP NetWeaver Product Management
July 2008
Agenda

Introduction to TREX in SAP NetWeaver

TREX Functions and Features

TREX Architecture and Details

TREX Plattform, Sizing Guidelines, ...

Summary
TREX is the one search technology in SAP solutions
TREX is deployed in over a dozen SAP products
TREX searches and analyses as well unstructured documents as structured business data
TREX in knowledge management provides search access to an extensible number of document repositories
TREX will provide the backend technology for Enterprise Search
Current use of TREX in SAP Solutions / Components

In SAP NetWeaver
- SAP NetWeaver Enterprise Search
- SAP Enterprise Portal
  - Knowledge Management (KM) platform
- SAP Business Intelligence (attachment, data aggregation in future)
- SAP KW (Training + Documentation solution)
- SAP Records Management

Further SAP solutions / components
- mySAP HR Expert Finder, e-Recruiting, Learning Solution
- mySAP PLM DMS (Document Management System)
- IS Automotive Vehicle Finder
- mySAP CRM
  - Internet Sales (Catalog Engine)
  - IC Web Client
  - Segment Builder
- ...more
TREX is a service provider for SAP solutions

UI
- Solution-specific UI

Apps
- IC Web Client

Engine
- TREX Engine

SAP Enterprise Portal UI (iViews)
- Intergrated Content in SAP EP

KM with SAP NetWeaver
- Shared TREX usage of Solutions

mySAP PLM DMS
- n - x TREX Services used

Index
- n TREX Services offered

- Solution specific TREX
TREX – Search Services offered

Search in
- Unstructured data (documents)
- Structured data (business objects)
- Full text
- Attributes

Different search modes
- Exact
- Linguistic: stemming, etc.
- Fuzzy: Search error tolerant
- Wildcards and truncations ( * or ? )
- Phrase search for complex expressions
- Boolean operators (AND, OR, NOT…)
- Highlighting / HTML conversion
- Content Snippets (Abstracts)
- Federated search
- …
Document Feature extraction
  ▶ Find characteristic keywords (noun phrases)

Find similar documents
  ▶ Based on document features

Document classification
  ▶ Assign a document to predefined categories

Term search
  ▶ Find better search terms; discover interesting relationships

Document clustering
  ▶ Discover sets of related documents

Term clustering
  ▶ Discover sets of related terms in the current corpus
Search for all types of attribute
- String, integer, floating point, date, and so on

Sort query results by any attribute
- For example, sort documents by date or by author

Support range search
- For example, find sales orders from the last two months

New:
- Multihost enabled
- Attribute search enabled not only for case-insensitive ASCII but also for case-insensitive Unicode
Additional Services – Example: Interactive Search

1. User enters search keyword Berlin

2. Search results: Hits grouped by attributes (overlapping) and listed by attribute value ranges (disjunct)

3. User clicks on group M-P in Street

4. Search results: Groups ordered by relevance for this search. All hits are in M-P streets in Berlin

5. User clicks street and sees hits
Agenda

Introduction to TREX in SAP NetWeaver

TREX Functions and Features

TREX Architecture and Details

TREX Plattform, Sizing Guidelines, ...

Summary
TREX APIs...

...may only be used SAP-internally

► Cannot directly be used by customers or partners

...can indirectly be accessed via other APIs

► e.g. the KM IMS API or the ABAP search engine service
TREX – Supported Languages

- Arabic
- Chinese trad.
- Chinese simpl.
- Czech
- Danish
- Dutch
- English
- Finnish
- French
- German
- Greek
- Hebrew
- Hungarian
- Italian
- Japanese
- Korean
- Polish
- Portuguese
- Norwegian bok.
- Norwegian nyn.
- Romanian
- Russian
- Spanish
- Swedish
- Thai
- Turkish
- ...more
Indexable MIME Types

- MS Word
- HTML
- XLS
- QuattroPro
- PDF
- Lotus Manuscript
- MS Rich Text

...approximately 200 text-containing file types
Agenda

Introduction to TREX in SAP NetWeaver

TREX Functions and Features

TREX Architecture and Details

TREX Plattform, Sizing Guidelines, ...

Summary
TREX 7.1 in next major release of SAP NetWeaver

Focus platform support on:
- Linux for x86_64
- Windows for x86_64

In Detail: http://service.sap.com/pam
Reasons for platform reduction

- Optimize scalability and performance for fewer platforms more efficiently
- Ensure a highly performance-optimized TREX for our customers
- Ensure completeness of 64-bit coding for next release
- No negative or limiting impact on other SAP solutions expected, because of TREX’ internal client/server architecture and planned appliance delivery
- Reduce cost of development and support and focus on available expertise at TREX development
Previous TREX releases upto 7.0

- Platform support of previous releases is of course valid and remains so until their end of maintenance.
- TREX 7.0 for SAP NetWeaver 2004s thus comes to 2014 in 5+2+1 model.

TREX releases in current use

- TREX 5.0 Out of maintenance
- TREX 6.0 End 2006
- TREX 6.1 2013
- TREX 7.0 2014

Current intention is to move 6.0 and 6.1 installations mostly to 7.0
Agenda

- Introduction to TREX in SAP NetWeaver
- TREX Functions and Features
- TREX Architecture and Details
- TREX Plattform, Sizing Guidelines, ...

Summary
Future TREX platform focus on Windows and Linux is a decision that has been made in relation to available development resources and expertises in TREX development.

It will enable more focussed development to optimize TREX performance and supportability.

It does not express any general platform preference trend at SAP.
Configuration and Administration

- TREX Search and Indexing
- Landscape Configuration
- Excursion: TREX Sizing
- RFC Connection
- Administration and Monitoring
TREX is Highly Scalable

- **TREX** can be distributed on multiple hosts
- **TREX** hosts can have dedicated roles (Indexing, searching, backup....)
- **TREX** processes can run multiple times within the same TREX instance on one host
- **TREX** hosts can be added any time
**TREX Hosts – Master and Slave**

- **Master Host**
  - Responsible for indexing
  - Can also be used for searching but not in default configuration
  - Manages original version of index

- **Slave Hosts**
  - Responsible for searching
  - Ensure performance during indexing times
  - Manages copy of master index
  - Index is created and updated using replication procedure
Replace Master Index Server and Queue Server if they become unavailable

Inactive if Master Server and Queue Server are available

Data has to be stored centrally

Use one backup server for whole system or one backup server per master server
TREX Hosts – Scalability

- Load Distribution for Searching and Indexing
- High availability for Searching
- Indexing larger data sets
- TREX is:
  - Scalable
  - Provides load balancing
  - Provides HA Solution for Search
Configuration and Administration

- TREX Search and Indexing
- Landscape Configuration
- Excursion: TREX Sizing
- RFC Connection
- Administration and Monitoring
An Approach to Sizing – Detailed Agenda

- KPIs for TREX sizing
- Quick information on BIA sizing
- Sizing Methods and Tools
  - Structured Data
  - Unstructured Data
- Example for document based TREX landscape
- Different landscapes for different stages
Key Performance Indicators for Sizing TREX

**CPU**
- Processing time: load during indexing and search
- Expressed in SAPS

**Disk**
- Storage of indexes and queues
- Expressed in MB

**Memory**
- Memory consumption during indexing and search
- Expressed in MB

**Network Load**
- Transferred amount of data
- KB per server request
Parameters Influencing TREX Sizing

- Amount of indexed data
- Search load
- Type of indexed data
- Number of languages
- Amount and frequency of delta indexing
- High availability needs
TREX Processing Structured and Unstructured Data

Object based Applications

Mainly Attributes

TREX Engine

Mainly text

Document based Applications

Indexes

BI Accelerator
Solution specific TREX

BI

TREX Engine

Indexes
Possible Sizing Methods and Tools

Rule of thumb
- “A typical CPU can process 4000 scenarios”

T-Shirt Sizing
- Simple algorithms with many assumptions

Formulas
- Simple or more complex

Offline Questionnaires
- For structured questions

Quick Sizer
- Based on users and throughput
Sizing TREX for Structured Data – Rule of Thumb

#attributes (mixed set of integer, string, text) x #values x #objects

< 100 million attributes per index server

Rule of Thumb

100 million attributes → about 1000 SAPS (2 GB RAM)
200 million attributes → about 2000 SAPS (4 GB RAM)

Varies largely depending on amount of
- String and text attributes
- Multivalue attribute
1. Use questionnaire to get an overview of your scenario
2. Use given formula to get a rough idea how many indexservers you need
3. Do hands on sizing by either
   - Using testdata from the application
   - Generating testdata on TREX machine, if you know datasets
4. Test indexing and search performance by monitoring CPU load and RAM consumption
5. Come to conclusion if your datasets allow larger amounts of attribute sets per indexserver or smaller ones
6. Split index or design landscape with different indexes
Assumptions

- 80% mixture of predominantly office documents
- 20% PDF HTML and ASCII
- Data volume of indexed content: 100GB

Leads to

- Compression ratio of 1:40 from size of source data (documents) to index size in main memory
- Searching: Up to 18 000 per hour 2000 SAPS / 6 GB RAM
- Indexing: 24 hrs time consumption 4000 SAPS / 20 GB RAM
## Required Disk Space – rule of thumb

<table>
<thead>
<tr>
<th></th>
<th>HTML/text Documents</th>
<th>Mixed set of Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index size + queue</td>
<td>Document set size x 2</td>
<td>Document set size x 0.5</td>
</tr>
<tr>
<td>(permanent)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Index snapshot size</td>
<td>(Document set size x 2) x 0.7</td>
<td>(Document set size x 0.5) x 0.7</td>
</tr>
<tr>
<td>(permanent)</td>
<td><em>in distributed scenarios without central storage</em></td>
<td></td>
</tr>
<tr>
<td>Temporary disk space</td>
<td>Document set size x 1.5</td>
<td>Document set size x 0.5</td>
</tr>
</tbody>
</table>

\[ \text{(Document set size x 3.5)} \]
\[ \text{(Document set size x 4.9)} \]
\[ \text{(Document set size x 1)} \]
\[ \text{(Document set size x 1.35)} \]
## Required space – rule of thumb for a Mixed Set of Documents

**Example: 50 GB of office and html/text documents**

<table>
<thead>
<tr>
<th></th>
<th>Disk Space</th>
<th>Main Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index size + queue (permanent)</td>
<td>Document set size x 0.5</td>
<td>Compression ratio 1:40</td>
</tr>
<tr>
<td></td>
<td><strong>25 GB</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>(50 GB x 0.5)</em></td>
<td></td>
</tr>
<tr>
<td>Index snapshot size (permanent)</td>
<td>Document set size x 0.5 x 0.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>17.5 GB</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>(50 GB x 0.5 x 0.7)</em></td>
<td></td>
</tr>
<tr>
<td>Temporary disk space</td>
<td>Document set size x 0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>25 GB</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>(50 GB x 0.5)</em></td>
<td></td>
</tr>
</tbody>
</table>

**Total:** 67.5 GB

**Main Memory:**

- Temporary disk space: 1.25 GB
An Example 1: Large Document Based TREX Landscape

2 Sets of Documents

Notes
- Basic quantity: 1.6 million / 800,000 documents per language
- Languages: 3 (E/G/J)
- Growth: 2000 per day new or changed

Discussion threads
- Basic quantity: 4.2 million
- Languages: 40
- Growth: 20,000 per day new or changed

Search requests
- 150,000 per day

Search requests
- 20,000 per day
An Example 2: Component Information System

Application System

TREX Engine

Notes Index

Discussions Index

1.6 million
about 30 GB

discussion threads

4.2 million
about 100 GB
An Example 3: Master Hosts

Master Host
mytrexmaster01

Master Host
mytrexmaster02

Logical index

discussions Index
An Example 4: Slave Hosts

- 2 slave hosts supporting master host system
- All in all 5 slave host systems (10 servers)
An Example 5: Discussion Threads Servers

- **master** hosts01+02
- **slave** hosts01+02
- **slave** hosts03+04
- **slave** hosts05+06
- **slave** hosts07+08
- **slave** hosts09+10

**Physical index sources**

- SAPS: 24,000
- RAM: 48 GB
- Disk: 60 GB
- Index is updated two times a day with about 10,000 new or changed documents per index run
- Running on 12 blades
An Example 6: Notes

- Using a delta index
- Merges every hour
- Index size:
  - German: 7GB
  - English: 8GB
  - Japanese: 1GB
An Example 7: Landscape Solution

About 35,000 users in different time zones
## An Example 8: Summary

- 5.8 million objects
- More than 20,000 new or changed documents per day
- 35,000 users
- 170,000 search requests per day
- 25 languages to be processed

### System Requirements

- **CPU:** 36,000 SAPS
- **RAM:** 48 GB
- **Disk:** 200 GB
  - 80 GB index size
  - 120 GB temporary space for index update

- Index updates twice a day for discussion threads and every hour for notes
TREX in Different Stages – Two Examples

<table>
<thead>
<tr>
<th>Stage I</th>
<th>Stage II</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initial Indexing</strong></td>
<td><strong>Initial Indexing</strong></td>
</tr>
<tr>
<td>- High load during initial indexing stage</td>
<td>- Less indexing and preprocessing required</td>
</tr>
<tr>
<td>- Multiple Master Indexservers and preprocessors to speed up initial indexing</td>
<td>- Use Master host (indexing) Slave host (searching) Concept</td>
</tr>
<tr>
<td></td>
<td>- Remove one host from landscape</td>
</tr>
<tr>
<td><strong>Adding more applications or content</strong></td>
<td><strong>Adding more applications or content</strong></td>
</tr>
<tr>
<td>- Start with small installation</td>
<td>- Add Master and/or Slave hosts</td>
</tr>
<tr>
<td>- One host for indexing and searching due to little update frequency and search requests</td>
<td>- More search load than expected and/or backup server necessary</td>
</tr>
</tbody>
</table>
No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP AG. The information contained herein may be changed without prior notice.

Some software products marketed by SAP AG and its distributors contain proprietary software components of other software vendors.

SAP, R/3, xApps, xApp, SAP NetWeaver, Duet, SAP Business ByDesign, ByDesign, PartnerEdge and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG in Germany and in several other countries all over the world. All other product and service names mentioned and associated logos displayed are the trademarks of their respective companies. Data contained in this document serves informational purposes only. National product specifications may vary.

The information in this document is proprietary to SAP. This document is a preliminary version and not subject to your license agreement or any other agreement with SAP. This document contains only intended strategies, developments, and functionalities of the SAP® product and is not intended to be binding upon SAP to any particular course of business, product strategy, and/or development. SAP assumes no responsibility for errors or omissions in this document. SAP does not warrant the accuracy or completeness of the information, text, graphics, links, or other items contained within this material. This document is provided without a warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability, fitness for a particular purpose, or non-infringement.

SAP shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials. This limitation shall not apply in cases of intent or gross negligence.

The statutory liability for personal injury and defective products is not affected. SAP has no control over the information that you may access through the use of hot links contained in these materials and does not endorse your use of third-party Web pages nor provide any warranty whatsoever relating to third-party Web pages.

Weitergabe und Vervielfältigung dieser Publikation oder von Teilen daraus sind, zu welchem Zweck und in welcher Form auch immer, ohne die ausdrückliche schriftliche Genehmigung durch SAP AG nicht gestattet. In dieser Publikation enthaltene Informationen können ohne vorherige Ankündigung geändert werden.

Einige von der SAP AG und deren Vertriebspartnern vertriebene Softwareprodukte können Softwarekomponenten umfassen, die Eigentum anderer Softwarehersteller sind.

SAP, R/3, xApps, xApp, SAP NetWeaver, Duet, SAP Business ByDesign, ByDesign, PartnerEdge und andere in diesem Dokument erwähnte SAP-Produkte und Services sowie die dazugehörigen Logos sind Marken oder eingetragene Marken der SAP AG in Deutschland und in mehreren anderen Ländern weltweit. Alle anderen in diesem Dokument erwähnten Namen von Produkten und Services sowie die damit verbundenen Firmenlogos sind Marken der jeweiligen Unternehmen. Die Angaben im Text sind unverbindlich und dienen lediglich zu Informationszwecken. Produkte können länderspezifische Unterschiede aufweisen.


SAP übernimmt keine Haftung für Schäden jeglicher Art, einschließlich und ohne Einschränkung für direkte, spezielle, indirekte oder Folgeschäden im Zusammenhang mit der Verwendung dieser Unterlagen. Diese Einschränkung gilt nicht bei Vorsatz oder grober Fahrlässigkeit.

Die gesetzliche Haftung bei Personenschäden oder die Produkthaftung bleibt unberührt. Die Informationen, auf die Sie möglicherweise über die in diesem Material enthaltenen Hotlinks zugreifen, unterliegen nicht dem Einfluss von SAP, und SAP unterstützt nicht die Nutzung von Internetseiten Dritter durch Sie und gibt keinerlei Gewährleistungen oder Zusagen über Internetseiten Dritter ab.

Alle Rechte vorbehalten.