Introduction to KRISTAL-IRMS

Overview
- Introduction to KRISTAL-IRMS
- Background
- Features of KRISTAL-IRMS
- Applications
- Further Development Plans
- Installing KRISTAL-IRMS

Information Retrieval
- Static Text Collection
- Inverted File (Index)
- Boolean Retrieval

However,
- Some documents are modified.
- New documents are created.
- Some documents are deleted.

KRISTAL-IRMS
- Knowledge Retrieval In Science & Technology Affiliated Literatures

Information Retrieval & Management System (IRMS) that combines the functions of an information retrieval system and a DB Management system (DBMS) developed by KISTI.

- FULL functions of an Information Retrieval System,
- SUBSET of data management functions of a DBMS, and
- DOCUMENT MANAGEMENT & SERVICE without DBMS software.

DB+IR

Full IRS
- High-speed/Large-scale full-text retrieval
- High-speed document indexing
- High-speed on-line document insert/delete/update
- High-speed document loading

Partial DBMS
**KRISTAL-IRMS History**

- **KRISTAL-I**
  - 1991. 5 - 1996. 2 (Information Retrieval using BASIS+)
- **KRISTAL-II**
  - 1996. 03 – (Information Retrieval Engine)
- **KRISTAL-2000**
  - 2000. 03 – (Information Retrieval & Management System)
- **KRISTAL-2002**
  - 2002. 10 – (Information Retrieval & Management System)
- **KRISTAL-IRMS**
  - 2006. 01 – (Information Retrieval & Management System) : commercial product level

**Background (1/2)**

- **Motives for Development**
  - Information Technologies based on native language/culture
    - KRISTAL started with indexing and retrieval technologies for Korean and Chinese texts.
    - Asian languages differ from Westerns in the respect of language processing technologies as well.
  - Complicated & Inefficient Information Management Systems
    - Prevailed document management systems are based on application-based loose coupling of DBMS and IRS.
    - In document management & service systems, IRS is used for text retrieval and DBMS for document management. Applications are used to couple these two separate software packages.
    - These systems uses only a small subset of DBMS features to store and manage documents.
    - If this small subset of management functions is implemented on IRS, document management systems can be very simple since it can be implemented based on IRS only, without expensive DBMS.

**Current Trends in Document Management Systems**

<table>
<thead>
<tr>
<th>Systems</th>
<th>Examples</th>
<th>Merits</th>
<th>Dements</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application-based Loose Coupling of DBMS and IRS</td>
<td>DBMS for document storage system and IRS for text retrieval. Additional applications connect DBMS and IRS for text and index sync.</td>
<td>DBMS is strong for management and IRS is strong for text retrieval.</td>
<td>Two separate software packages (EXPENSIVE), Complex system architecture, Gap between a document (DBMS) and its index (IRS).</td>
<td>Prevalied</td>
</tr>
<tr>
<td>Loose coupling of IR features on DBMS system</td>
<td>Oracle Cartridge, IBM DB2’s Text Extender</td>
<td>Extendable data types</td>
<td>Overhead for extended data types, SQL-based retrieval only, Slow search speed.</td>
<td>Almost Dead</td>
</tr>
<tr>
<td>Tight coupling of IR features on DBMS system</td>
<td>Odysseus (KAIST)</td>
<td>Simple document management</td>
<td>Slow search, Weak retrieval features.</td>
<td>Limited</td>
</tr>
<tr>
<td>Tight coupling of DBMS features on IR system</td>
<td>KRISTAL IRMS</td>
<td>Fast Search, Full IR features, Online data management</td>
<td>Still weak in document management in the respect of DB manager.</td>
<td>Will be prevailed?</td>
</tr>
</tbody>
</table>

**Background (2/2)**

- Management Application
- Retrieval Application
- Database Manager
- Users

(a) DBMS-IRS Coupling Architecture

(b) IRMS Architecture
Strategic Focus on KRISTAL Development

- Focus on high-tech information retrieval and service technologies
- Develop an extendible IRMS that combines a search engine and a DBMS
- Reflect requirements from IRMS-based information service systems

Features of KRISTAL-IRMS

- Loading large scale data at a high speed
- Internationalize through Unicode
- Natively handle data
- XML

KRISTAL Features(1)

- Document Storage and Management
  - Fast uploading of large-scale data
  - Stable structure not affected by the size of the document or DB
  - Unicode-based documents and index storage
  - XML storage and management
  - Support various types of data (Text, Multimedia, BLOB...)

- Database Management
  - Guarantee online data management through transaction processing
  - Provide a Primary Key for redundancy checking and management
  - GUI DB management tool
  - Easy DB uploading and backup

KRISTAL Features(2)

- Retrieval System
  - Fast retrieval through multi-threaded database access
  - Concurrent query processing through process-pool method
  - High recall rate
  - Vector/Basic search model
  - Similar documents retrieval and Retrieval in results

- Indexing System
  - Provide various types of indexing such as word character-based indexing, morpheme and yes indexing, compound noun extendable query processing
  - Apply a Korean Morphological analyzer developed by KISTI
  - Unicode-base

- Application System
  - Provide various types of libraries required for developing clients
  - C/C++, JAVA APIs to access KRISTAL servers
  - Various indexing Method
  - Fast and accurate built-in Morphological analyzer
  - Unicode-based indexing
Areas of Applications

Applications 1/5: Bibliography Retrieval

- Retrieval System for S&T Literatures of KISTI
  - URL: http://www.yeskisti.net
  - About 50 million plain documents in Korean and/or English language
  - Indexing Korean Texts
    - Korean Morpheme Analyzer
      - Ex) "가 잘 먹어요." → "가공. 잘먹어요.
  - Indexing English Texts
    - Token-level Indexing
      - Ex) "traveling to Vietnam" → "TRAVELING", "TO", "VIETNAM"
      - Optionally stopwords (such as "TO") can be removed
      - Optional stemming can be applied
    - Raking Retrieval Model (Vector)

Applications 2/5: Historic Articles of Korea

- Korean History On-line
  - URL: http://www.history.go.kr
  - About 5 million XML documents in Chinese and/or Korean letters
  - Indexing Chinese Letters
    - Character-level Indexing
    - Phonetic Value Indexing
    - Bi-gram Indexing
    - Ex) "가져다 줄려" → "가져다", "줄려"
    - With many other techniques to deal with Chinese letters in Korean historic literatures.
  - Boolean Retrieval Model

Applications 3/5: Encyclopedia for Local Areas

- Digital Encyclopedia of Seongnam City, Korea
  - URL: http://seongnam.grandculture.net
  - About 5 thousand XML documents in Chinese and Korean letters. Every personal name, place name, historic event is tagged.
  - Management & Service is synchronized with KRISTAL-IRMS.
    - Local citizen can post his/her own article to the encyclopedia.
    - Professional writers can reflect the citizen’s opinion to their article in real time.
    - Knowledge can be circulated to higher quality.
  - Boolean Retrieval Model

Articles by Professionals

Articles by citizens
Applications 4/5: Scientific Data Analysis

- Protein Sequence Analysis
  - URL: http://proses.kisti.re.kr
  - About 100 thousand of protein sequences
  - Subcellular location(s) for a new protein sequence can be predicted.
  - Indexing Protein Sequences
    - Overlapped Pentagram
      - EJU, ACDEFGH \( \rightarrow \) "ACDEF", "CDEFG", "DEFGH", "EFGH"
    - Automated Text Categorization

Applications 5/5: Other Sites (1/2)

- Scientific & Technical Information Services of KISTI
  - http://techtrend.kisti.re.kr (Technical Trends Database)
  - http://next10.yeskisti.net (Next Generation Technology Information Service)
  - http://www.nktech.net (S&T Information of North Korea)
- Full Text Search of Korean Books
  - http://www.booktopia.com (Booktopia)
- News Photo Management Systems
  - Korean Economy Daily, Kookmin Ilbo, etc. (For Intranet)

Applications 5/5: Other Sites (2/2)

- Retrieval Systems for Historical Literature
  - http://sjw.history.go.kr (Seung-Jeong-Won Diary)
  - http://e-kyujanggak.snu.ac.kr (Kyu-Jang-Gak)
  - http://www.minchu.or.kr (Korean Classics Research Institute)
- Retrieval System for Scientific Information
- Photo Album with Full Text Search
  - http://www.animalpicturesarchive.com
- Photo Album with Full Text Search
  - http://www.animalpicturesarchive.com
- And many more will be on-line sooner or later.

Further Development Plans

- Support KNOWLEDGE CIRCULATION in Asian language texts
- Support SCIENTIFIC DATA ANALYSIS using data mining
- Do not need to buy an expensive RDBMS for document management
- Asian Language Processing / Scientific Data Processing
- SQL-like IMQL (Information Management Query Language)
- Efficient Offline/Online Data Management
- Distributed Information Management & Retrieval
- Improvement of User Supporting Tools
Installing KRISTAL-IRMS (1/3)

• KRISTAL-IRMS System Requirements
  - OS: Linux (Complete Installation recommended)
  - Other UNIX platforms such as Solaris and HP-UX are also supported under limited conditions.
  - 512MB of RAM (Recommended 1GB or more)
  - GCC 3.x or 4.x with various development tools provided by Linux Distributions.

• Downloading KRISTAL-IRMS
  - [http://www.kristalinfo.com/download/#kristal](http://www.kristalinfo.com/download/#kristal)
  - Download KRISTAL-2002.2.1.1.tar.gz and save to an appropriate directory.

Installing KRISTAL-IRMS (2/3)

• Installation
  - Restore source files from the tar archive
    - tar zxf KRISTAL-2002.2.1.1.tar.gz
  - Compile
    - cd KRISTAL-2002.2.1.1
    - sh INSTALL.sh
      - This will take about 20 minutes or more depending on the specification of the machine.
    - cd ..
    - ln -s KRISTAL-2002.2.1.1 KRISTAL
      - If the current directory is "/home/kristal",
        $KRISTAL_HOME can be shortened to "/home/kristal/KRISTAL".
  - Add $KRISTAL_HOME/bin to your path

• Directories
  - $KRISTAL_HOME/bin: daemon, loader, dumpers, etc.
  - $KRISTAL_HOME/lib: dictionaries and C++ libraries
  - $KRISTAL_HOME/include: KRISTAL headers

Installing KRISTAL-IRMS (3/3)

• Directories and Files

<table>
<thead>
<tr>
<th>Directory</th>
<th>Files</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>bin</td>
<td>kristald</td>
<td>KRISTAL daemon</td>
</tr>
<tr>
<td></td>
<td>kristal_dbadmin</td>
<td>Off-line database management tool</td>
</tr>
<tr>
<td></td>
<td>kristald_stop</td>
<td>Stops KRISTAL daemon</td>
</tr>
<tr>
<td>lib</td>
<td>libcom.a</td>
<td>Communication library</td>
</tr>
<tr>
<td></td>
<td>libclient.a</td>
<td>Client API library</td>
</tr>
<tr>
<td></td>
<td>libconv.a</td>
<td>Iconv library</td>
</tr>
<tr>
<td></td>
<td>libshare.a</td>
<td>Various shares</td>
</tr>
<tr>
<td></td>
<td>libidx.a</td>
<td>KRISTAL Indexer library</td>
</tr>
<tr>
<td></td>
<td>libxerces-c.a</td>
<td>XML parser (XERCES) library</td>
</tr>
<tr>
<td></td>
<td>k_dic/*</td>
<td>Dictionaries for Korean Morpheme Analysis</td>
</tr>
<tr>
<td>include</td>
<td>client/*.h</td>
<td>KRISTAL Header files for C++ client</td>
</tr>
</tbody>
</table>

Indexing English and Korean

- Token level indexing is sufficient.
- Stemming or stopword removal can be applied.

English: "My son goes to Elementary School.
Korean: "³ªÀÇ -¾ÆµéÀº -ÃʵîÇб³¿¡ -°£´Ù ."
Uzbek: ???

- A Hangeul token usually consists of NOUN + POSTFIX.
- Token is not sufficient for indexing natural Korean texts.
- Korean Morpheme Analysis should be applied to extract index terms.
- Complex noun should be separated to basic nouns.
Thank you for your attention!

http://www.kristalinfo.com