Software benchmarking in digital preservation

Kresimir Duretec, Artur Kulmukhametov, Andreas Rauber and Christoph Becker
Vienna University of Technology and University of Toronto - BenchmarkDP project

Carl Wilson
Ope Preservation Foundation

Bengt Neiss
National Library of Sweden - PREFORMA project

February 11, 2016
Benchmarking initiatives in the Information Retrieval field
A Range of Benchmarking Initiatives

- MIREX: music
- MSD: music
- CHEMTREC: chemical patents
- LifeCLEF: images/video/audio (bird songs)
- Many more text-based… (e.g. Reuters, ClueWeb, Govdocs, …)
Welcome to MIREX 2015

This is the main page for the eleventh running of the Music Information Retrieval Evaluation eXchange (MIREX 2015). The International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) at the Graduate School of Library and Information Science (GSLIS), University of Illinois at Urbana-Champaign (UIUC) is the principal organizer of MIREX 2015.

The MIREX 2015 community will hold its annual meeting as part of The 16th International Conference on Music Information Retrieval (ISMIR 2015), which will be held in Malaga, Spain, October 26th-30th, 2015. The MIREX plenary and poster sessions will be held during the conference.

J. Stephen Downie
Director, IMIRSEL

Task Leadership Model

Like ISMIR 2014, we are prepared to improve the distribution of tasks for the upcoming MIREX 2015. To do so, we really need leaders to help us organize and run each task.

To volunteer to lead a task, please add your name to the “Captains” column on the new 2015 Task Captains page. Please direct any communication to the EvalFest mailing list.

What does it mean to lead a task?

- Update wiki pages as needed
- Communicate with submitters and troubleshooting submissions
Evaluation in Music IR: History

- Discussion started at ISMIR 2001
  - evaluation frameworks
  - standardized test collections
  - tasks and evaluation metrics
- IMIRSEL project started 2002:
  (International Music Information Retrieval Systems Evaluation Laboratory), Univ. of Illinois, Stephen Downie
- First Audio Description contest at ISMIR 2004
- MIREX (Music Information Retrieval Exchange) started in 2005
- Annual, in connection with ISMIR conferences
- Evaluating many approaches of the MIR domain
2004 Audio Description Contest

- First attempt towards comparative benchmarking of MIR algorithms
- Five different tasks
  - Genre Classification
  - Artist Identification
  - Melody Extraction
  - Tempo induction
  - Rhythm Classification
- Some training/test data made available to participants
- Automatic evaluation
- Test for robustness of algorithms
MIREX 2012 Tasks

- Audio Classification
  - Audio US Pop Genre Classification
  - Audio Latin Genre Classification
  - Audio Music Mood Classification
  - Audio Classical Composer Identification
- Audio Cover Song Identification
- Audio Tag Classification
- Audio Music Similarity and Retrieval
- Symbolic Melodic Similarity
- Audio Onset Detection

- Audio Key Detection
- Real-time Audio to Score Alignment
- Query by Singing/Humming
- Audio Melody Extraction
- Multiple Fundamental Frequency Estimation & Tracking
- Audio Chord Estimation
- Query by Tapping
- Audio Beat Tracking
- Structural Segmentation
- Audio Tempo Estimation
MIREX 2015 Tasks

- Grand Challenge on User Experience
- Audio Classification (Train/Test) Tasks
  - Audio US Pop Genre Classification
  - Audio Latin Genre Classification
  - Audio Music Mood Classification
  - Audio Classical Composer Identification
  - Audio K-POP Mood Classification
  - Audio K-POP Genre Classification
- Audio Cover Song Identification
- Audio Tag Classification
- Audio Music Similarity and Retrieval
- Symbolic Melodic Similarity
- Audio Onset Detection
- Audio Key Detection
- Real-time Audio to Score Alignment (a.k.a Score Following)
- Query by Singing/Humming
- Audio Melody Extraction
- Multiple Fundamental Frequency Est./Tr.
- Audio Chord Estimation
- Query by Tapping
- Audio Beat Tracking
- Structural Segmentation
- Audio Tempo Estimation
- Discovery of Repeated Themes & Sec.
- Audio Downbeat Estimation
- Audio Fingerprinting
- Singing Voice Separation
- Set List Identification
- Music/speech classification and detection
Classification Tasks

- Fully automatic evaluation, objective -> relevant?
- For stable results, cross-validation is used:
  - full data set is split into n folds
  - in n iterations, n-1 parts of the data set are used for training the algorithm (learning), the remaining part is used for testing
  - final result is average performance of n folds

- In publications, 10 fold CV, in MIREX typically 3-fold CV
- Significance tests performed
  (or standard deviation of folds given)
- Comparable results!
Audio Music Similarity and Retrieval Task

- Similarity retrieval rather than classification
- Evaluated by human judgements: human listening tests (first @ MIREX 2006)
- Evalutron 6000: http://www.music-ir.org/evaluation/eval6000
- Test of statistical significance: Friedman test
MIREX Human Evaluation

- 60 randomly selected queries
- ~20 human evaluators
- 7-8 ranked lists per evaluator
- 3 evaluations per ranked list
- 2 evaluation scales:
  - broad scale: very/somewhat/not similar
  - fine scale: between 0 and 10 (10 = best)
MIREX Evalutron 6000

Audio Music Similarity & Retrieval 2006

Welcome kuku1 Sign out Change My Settings

Home Audio Player Selection My Assignment Instructions

THIS PAGE CONTAINS 25 CANDIDATES FOR QUERY ID # 6

Query ID#6
Listen to Candidate #b005105

Select Broad Category
- NOT Similar
- Somewhat Similar
- VERY Similar

Select Fine Score

Listen to Candidate #b008631

Listen to Candidate #b002824

Select Broad Category
- NOT Similar
- Somewhat Similar
- VERY Similar

Select Fine Score

Align Player

First Mid Last

First Mid Last

First Mid Last

FACULTY OF INFORMATICS
MIREX2006:
MusicSim: Metadata Statistics - Results

- Results on the top 20 most similar
Challenges for Music Retrieval Benchmarking

- **Data - and access to it**
  - real-world data set (- but how to get & use legally?)
  - sufficient (?) size
  - sufficient quality

- **Metadata**
  - high-quality labels (production-style)
  - ground truth annotation (can be very very time-consuming!!)

- **Evaluation**
  - automatic vs. human evaluation
  - which are the proper evaluation measures?
    - how to perform tests properly?
MIREX Summary

- growing number of tasks
- growing number of people interested
- growing size of data sets
- data set issues remain (copyright, distribution, insufficient)

- MIREX is open to everyone interested in evaluation
- democratic process via mailing list and Wiki
MIREX Summary

Benchmarking needs

- an infrastructure
- Staff
- Participants agreeing on a specific task
  - representative?
  - relevant?
- Agreement on evaluation measure
- Benchmarking vs. Competition

Collaboration & Money
Software benchmarking in Digital Preservation
Software tools in digital preservation

- many tools (>400 tools in COPTR registry http://coptr.digipres.org)

- categories
  - identification, validation & characterisation
  - migration
  - emulation
  - ....

- high quality required
# Software benchmarks - definition

<table>
<thead>
<tr>
<th>Source</th>
<th>Definition</th>
<th>key words</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEC glossary</td>
<td>“a &quot;benchmark&quot; is a test, or set of tests, designed to compare the performance of one computer system against the performance of others”</td>
<td>test, compare, performance</td>
</tr>
<tr>
<td>IEEE Glossary</td>
<td>“1. a standard against which measurements or comparisons can be made. 2. a procedure, problem, or test that can be used to compare systems or components to each other. 3. a recovery file”</td>
<td>measurements, comparison, recoverable file</td>
</tr>
<tr>
<td>W. Tichy</td>
<td>“a benchmark is a task domain sample executed by a software tool. A human and computer records well-defined performance measurements. A benchmark provides a level playing field for competing ideas, and (assuming the benchmark is sufficiently representative) allows repeatable and objective comparisons”</td>
<td>task domain sample, performance measurements, repeatable objective comparison</td>
</tr>
<tr>
<td>Sim et al.</td>
<td>“a test or set of tests used to compare the performance of alternative tools or techniques”</td>
<td>test, compare</td>
</tr>
</tbody>
</table>

A DP Software Benchmark is a set of tests which enable rigorous comparison of software solutions for a specific task in digital preservation.
Benchmarks in other fields

- specifications
- tools & methods that need benchmarks
- community
- yearly meeting
- iterative process
- results

TU VIENNA
FACULTY OF INFORMATICS
Motivating comparison

- Digital preservation
- Function
- Dataset
- Ground truth
- Performance measures

- What is a benchmark supposed to compare?
- Which quality aspects are addressed?
- What would be the benefits?
Motivating comparison

- Digital preservation
- Motivating comparison
- Function
- Dataset
- Ground truth
- Performance measures

- Software quality
  - Many aspects
- We need a model
  - ISO SQUARE
  - Correctness
  - Reliability
  - Performance efficiency
Digital preservation

Motivating comparison

**Function**
- Specific task tool is expected to perform
- Representative of problems in actual practice
  - Identification
  - Validation
  - Characterization
  - Migration

Dataset

Ground truth

Performance measures
Dataset

- **Digital preservation**
- **Motivating comparison**
- **Function**
- **Dataset**
  - set of digital objects
    - images
    - audio
    - video
  - relevant
  - accurate
  - complete
  - How big the dataset should be?
  - How can it be generated?
- **Ground truth**
- **Performance measures**
Ground truth

Digital preservation
Motivating comparison
Function
Dataset
Ground truth
Performance measures

- correct answers
- optional element
- machine readable

Which elements should it contain?
Performance measures

Digital preservation

Motivating comparison

Function

Dataset

Ground truth

Performance measures

- fitness of the benchmarked tool
- type
  - quantitative
  - qualitative
- calculated
  - by human
  - by machine
Conclusion - IPRES 2015
Conclusion - now
How can you help?

- For a start defining all five components can be too hard
- You can focus on a motivating comparison only
  - write a short “benchmark brief”
  - What is the task for which you would like to see a benchmark
    - no need for technical details
  - What would you like to see compared by the benchmark
    - correctness, speed, …
- Give some tool examples
- Publish your “benchmark brief”
  - OPF blog post might be a perfect place
  - we are currently still lacking central place to collect all the specifications and comments
Property value extraction from page-based documents in MS Word 2007
Page based documents

- Genres
  - books
  - articles
  - contracts
  - ....

- File formats
  - PDF
  - DOC
  - DOCX
  - ODT
  - ...

FACULTY OF INFORMATICS
Properties and significant properties

- file size = 12kB

- significant properties
Characterisation tools

- various characterisation tools
- capable of extracting property values
- which?
- are those values correct?
Motivating comparison

- Coverage
  - Compare coverage of needed properties
- Correctness
  - Is the tool correctly extracting the values of given properties
# Function

## Extract values of various properties

<table>
<thead>
<tr>
<th>Property name</th>
<th>Explanation</th>
<th>How is it measured</th>
<th>Value type</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>numberOfPage</code></td>
<td>number of pages in a document</td>
<td>Count the number of pages in a document.</td>
<td>integer</td>
</tr>
<tr>
<td><code>textValue</code></td>
<td>free text in the document</td>
<td>Extract text that is not part of any other element (table, figure, header, footer) from the document. Any other formatting like extra spaces tab values should be ignored. If included in the returned output these will still be ignored by the tool doing comparison.</td>
<td>string</td>
</tr>
<tr>
<td><code>numberOfFigure</code></td>
<td>number of figures in a document</td>
<td>Count the number of figures in a document.</td>
<td>integer</td>
</tr>
<tr>
<td><code>numberOfTable</code></td>
<td>number of tables in a document</td>
<td>Count the number of tables in a document.</td>
<td>integer</td>
</tr>
<tr>
<td><code>numberOfLine</code></td>
<td>number of lines of free text in a document</td>
<td>Count the number of lines in a free text in the document.</td>
<td>integer</td>
</tr>
</tbody>
</table>
# Performance measures

<table>
<thead>
<tr>
<th>Software Quality Measure</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purpose of the Software Quality Measure</strong></td>
<td>Evaluate the completeness of a tool in covering certain property by calculating the percentage of documents where a tool returned value for that property and it was supposed to return a value.</td>
</tr>
<tr>
<td><strong>Measurement Function</strong></td>
<td>$C_v = \frac{DP}{N}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Software Quality Measure</th>
<th>Correctness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purpose of the Software Quality Measure</strong></td>
<td>Evaluate the correctness of a tool in calculating certain property values by calculating the percentage of documents where a tool returned correct value for that property.</td>
</tr>
</tbody>
</table>
What are the critical points

- Function
  - Which properties are important?
  - How do we find those?
- Dataset
  - What should the dataset cover?
  - How do we create such dataset?
- Performance measures
  - Which measures are needed?
  - How are they calculated?
Photo migration benchmark
Introduction

- Raw photographs
- Various formats
  - obsolescence
- Migration to a standardised format
  - Adobe Digital Negative (DNG)
The problem

- Tools for photo migration

- Which tool does a migration correctly?
  - Can we trust them?
  - What is the evidence?

- How to select the best tool for migration to DNG?
Photo migration benchmark

- Raw photographs
  - made by digital cameras
  - many models -> save data in different file formats
- Motivating comparison
  - compare the correctness of migration from a RAW file format to DNG
- Function
  - migrate from RAW to DNG
- Dataset and ground truth
  - collection of photographs in RAW file formats
  - created by sampling real world collections
Photo migration benchmark

- Dataset identified
  - The photo collection from [http://www.rawsamples.ch/](http://www.rawsamples.ch/)
  - Sample photos from The Getty Research Institute

- Performance Measures
  - Metadata migration score (MMS)
    - asserts changes in metadata.
  - Content migration score (CMS)
    - calculates difference in content, based on SSIM.
  - Amount of failures (AF)
    - show how many times a tool failed during migration.