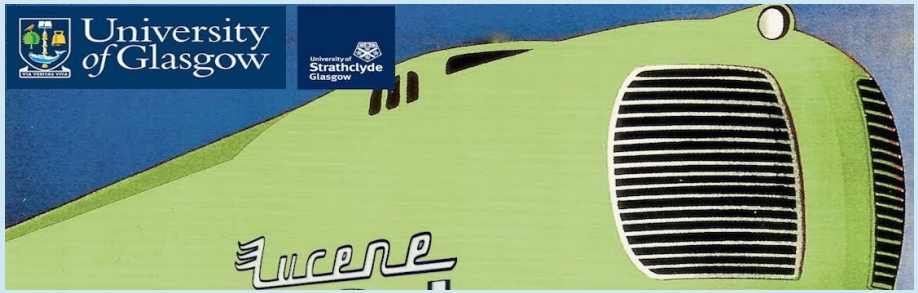
Using Lucene for Teaching and Learning IR: The University of Granada case of study

Juan M. Fernández-Luna
Universidad de Granada, Spain
jmfluna@decsai.ugr.es



Lucene4IR: Developing IR Evaluation Resources using Lucene

Outline

Introduction

SulaIR: a first attempt of Teaching and Learning IR application

Contextualization of current IR subjects @ UGR – Bologna Process

Lucene across IR subjects @ UGR

Needs for students and lecturers.

Outline

Introduction

SulaIR: a first attempt of Teaching and Learning IR application

Contextualization of current IR subjects @ UGR – Bologna Process

Lucene across IR subjects @ UGR

Needs for students and lecturers.

Introduction

@UGR → First IR-related subject at Master level.

- Introduction to IR in the context of Probabilistic Graphical Models.
- Whole IR process.
- No lab practice.

Bologna process → 3 subjects (1 bachelor, 2 masters).

- Lab work. How could we approach the practical work?
    - Writing code from the scratch?
    - Using any API (Lucen, Terrier, Lemur, MG,…)?

Introduction

Introduction

First choice: from the scratch.

But… we did a mistake:

- Programming problems – skills.
- Understanding the IR process and transferring to a programme.
- → We could not advance at good pace. Students did not acquire the learning outcomes.

Introduction

Then, we moved to an API: Lucene.

- Students just focused on learning an API and identifying the classes and methods that they need.
- Open source.
- Documentation, community, widely used.
- ...

Outline

Introduction

SulaIR: a first attempt of Teaching and Learning IR application

Contextualization of current IR subjects @ UGR – Bologna Process

Lucene across IR subjects @ UGR

Needs for students and lecturers.

SulaIR – TLIR Tool

- Teaching and learning innovation project @ UGR
- Developed during the academic course 2010/2011.
- Used in a IR-related master subject.
- SulaIR is an interactive tool that allows to visualize the complete IR process and interacts with the system in each stage in which such process can be decomposed (document preprocessing, indexing, querying, retrieval and relevance feedback).

SulaIR – TLIR Tool

Objectives of this tool:

1. To support the student in the learning task of the main concepts of IR by means of a easy and friendly interaction with an educational software.

2. To facilitate the lecturer her teaching IR labour, allowing the use of new technologies.

3. To provide a permanent platform, intuitive and simple, accessible from Internet, from any type of computer and operating system, and extensible with new modules.

SulaIR – TLIR Tool

Some features:

1. Designed an implemented from the scratch → Total autonomy to deal with data structures and UI integration → No Lucene.

2. Coded in Java as a desktop application. First version of web application.

3. It could process XML, TREC formats, HTML.

4. In its last version, a crawler was implemented for HTML documents.

5. The instructor and the student focused on the process not in the implementation details.

SulaIR – TLIR Tool

Demo

Outline

Introduction

SulaIR: a first attempt of Teaching and Learning IR application

Contextualization of current IR subjects @ UGR – Bologna Process

Lucene across IR subjects @ UGR

Needs for students and lecturers.

Contextualization of IR subjects

Bologna Process

5th year	Master in Data Science						**Information Retrieval and Recommender systems**				**Master thesis**
	Master in Computer Science						**Information Management in the Web**				**Master thesis**

4th year	**Bachelor thesis**	**Bachelor thesis** **Information Retrieval** **Digital Libraries**	**Bachelor thesis**	**Bachelor thesis**	**Bachelor thesis**
3rd year	Software Engineering	Information Systems	Computer Engineering	Intelligent Systems	Information Technologies
2nd year	Common subjects				
1st year	Common subjects				

Computer Science Degree

Outline

Introduction

SulaIR: a first attempt of Teaching and Learning IR application

Contextualization of current IR subjects @ UGR – Bologna Process

Lucene across IR subjects @ UGR

Needs for students and lecturers.

Lucene across IR subjects @UGR

Information Retrieval

(Degree in Computer Science, 4th year, 1st semester, 6 ECTS [3 lectures + 3 lab → 2h lectures + 2h lab / per weeek)

Lectures: Theory in IR foundations and "advanced topics" (clustering, classification,…).

Lab work: At the beginning, some IR tasks from the scratch, later working with Lucene → But the Million dollar question is:

do students have to know how to implement VSM or just using it for retrieval?

Our answer: the averaged computer scientist, only using it.

Now, only using Lucene API → Time and effort savings.

Lucene across IR subjects @UGR

Information Retrieval

Lab contents:

1. Development of a desktop search engine application.
    1.1. Text extraction: Tika.
    1.2. Document preprocessing: Lucene.
    1.3. Indexing: Lucene.
    1.4. Retrieval: Lucene.
2. Introduction to Solr.

Lucene across IR subjects @UGR

Information Retrieval - Findings

- Difficult start as student don't know too much Java. C++ a possibility but not too much documentation.

- Lucene is a monster → this provokes a certain rejection to the students.

- If they are left alone, fail with Lucene.

- We are having lots of problems with the different Lucene versions.

- Large amount of information about how to do basic tasks in Lucene but versions mixed → Mess for the students.

- Good bibliography not updated to the last versions. No documentation in Spanish.

- Luke very useful but not too much support.

- Everything goes relatively smooth until they have to work with fields and more complex queries.

Lucene across IR subjects @UGR

Information Management in the Web

(Master in Computer Science, 1st year, 4 ECTS [2 lec. + 2 lab])

Social networks analysis + Information Retrieval + Recommender systems

IR foundations in 3 weeks (1.5h lectures + 1.5h lab / per week).

Lab work:

1) Students without previous IR knowledge: Given the code of the basic class for indexing and searching tasks in Lucene, they have to write a simple search engine (indexing and retrieval apps). Luke for verification.

2) Students with previous IR knowledge: Implementation of a new IR model in Lucene or the same as 1) with Solr.

Lucene across IR subjects @UGR

Information Management in the Web

Our findings:

- Total guidance → the ABC of building a search engine step by step.

- Very practical for the students → they say no need to know more ;-)

Lucene across IR subjects @UGR

Information Retrieval and Recommender Systems

(Master in Data Science, 1st year, 2nd semester, 3 ECTS [1.5 lect. + 1.5 lab])

Information Retrieval + Recommender systems

IR foundations in 4 weeks + advanced topics related to data science (machine learning for IR) (1h theory + 1h practice / per week).

Lab work:

1) Preprocessing large amounts of texts with Tika and Lucene.

2) Indexing with Lucene and inspection with Luke.

Lucene across IR subjects @UGR

Information Retrieval and Recommender Systems

Our findings:

- In practical terms, they are not so interested in retrieval itself as in processing large amount of texts → Lucene seen as a tool for pre-processing text and to have it ready for further tasks.

- More interest on the Lucene scalability and distribution features.

- More interest on how they can use Lucene as the base for machine learning process.

- Different backgrounds (CS, Maths, IS, Statistics,...) → difficult programming. Maybe moving to R this year.

Lucene across IR subjects @UGR

IR-related Bachelor and Master thesis

(4th year of Computer Science Degree, Masters in Data Science and Computer Science - 2nd semester and 12 ECTS).

IR-based projects with Lucene/Solr as tools for much bigger developments.

Our findings:

- Too much problems for the students in order to find advanced documentation → too much trial and error.

- No support for "real" documents.

- They usually have to write their own evaluation tools from the scratch.

- Difficulties to develop new retrieval models or ranking functions.

Outline

Introduction

SulaIR: a first attempt of Teaching and Learning IR application

Contextualization of current IR subjects @ UGR – Bologna Process

Lucene across IR subjects @ UGR

Needs for students and lecturers.

Needs for students and lecturers

- Documentation:
    - Lucene tutorial/documentation for teaching and learning, updated to its last version.
    - Lucene tutorial/documentation for more advanced tasks.
- Software for supporting the teaching and learning process:
    - Lucene-based SulaIR: SulaIR-L.
    - Visual, command line and classes for evaluation tools (TREC-based experimentation).
    - Lucene native support for common types of documents.

Thank you!!