Proposition de Sujet de Mastère 2017-2018

Hybrid indexing tool for Arabic information retrieval system

Laboratories/research groups: LISI Computing Laboratory for Industrial Systems, INSAT, JARIR: Joint group for Artificial Reasoning and Information Retrieval (www.jarir.tn).

Head: Pr. Yahya Slimani, ISAMM, yahya.slimani@gmail.com

Supervisor: Dr. Ibrahim Bounhas, ISD, bounhas.ibrahim@gmail.com

Descritpion:

The goal of this project is to enhance an existing hybrid indexing tool in order to give better efficiency (**speed**, resources) and effectiveness (recall, precision,…). The JARIR group has been working on developing an Arabic IRS (Information Retrieval System) based on a hybrid index (Ben Guirat et al. 2016). The proposed approach is to build a multilevel index where the hierarchical structure represents the semantic relations between the different word forms (root, verbal pattern and stem). Given the existent tool, this project aims to:

1- Develop a performant hybrid indexing tool enhancing the capacity of the IRS. A tool developed by our group will be the starting point.
2- Integrate the proposed tool on Terrier[1] IR Platform using BM25 model.
3- Perform experiments based on a **large scale corpus** (Arabic newswire LDC test collection).
4- Using MADAMIRA[2] and Alkhalil[3] tools to add the lemma unit to the hybrid index.
5- Perform interpretations based on performance and significance tests using TANAGRA[4].

Key-words: Arabic Information Retrieval, Hybrid index, algorithm complexity
Languages: JAVA

Reference :

S. Ben Guirat, Bounhas, I., and Slimani, Y., “A Hybrid Model for Arabic Document Indexing”, in 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2016), May 30 - June 1, 2016 Shanghai China, 2016

[1] http://terrier.org/
[2] https://camel.abudhabi.nyu.edu/madamira/
[3] https://sourceforge.net/projects/alkhalil/
[4] http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html