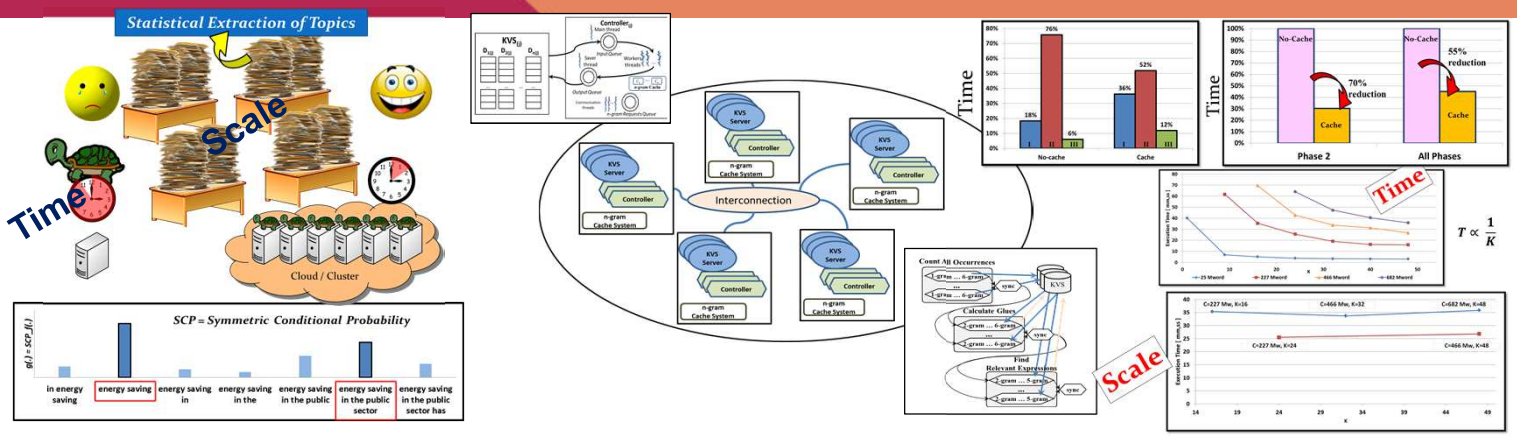
NOVALINCS
LABORATORY FOR COMPUTER SCIENCE AND INFORMATICS

ISEL
Instituto Superior de Engenharia de Lisboa

12th IEEE International Conference on eScience
October 24-26, 2016 – Baltimore, Maryland, USA

An *n*-gram cache for large-scale parallel extraction of multiword relevant expressions with LocalMaxs

Carlos Gonçalves - cgoncalves@deetc.isel.pt
Phd Student. ISEL/IPL – FCT/UNL
Teaches distributed systems. Researches parallel and distributed solutions to improve the performance of relevant expression extraction and document correlation algorithms.
Advisors:
- Professor José C. Cunha
- Professor Joaquim F. Silva



Motivation and Challenges

- Enable the extraction of relevant multiword expressions from very large natural language *corpora*, using statistical methods in acceptable time
- Use of parallel and distributed computing supported by local clusters and public clouds
- Multiword relevant expressions capture the core contents of document semantics. Only strong average cohesion (glue) among words points to multiword relevant expressions

Approaches

Sequential		Parallel & Distributed
Very time-consuming !!!	→	Parallel: To reduce Time
Huge memory-demanding !!!	→	Distributed: To fit in Memory

Methods and Techniques

- Generic architecture capable of:
  - Execute algorithms based on statistical *n*-gram models;
  - Being executed in cluster or cloud environments
- Phase 1 counts the *n*-gram occurrences
  - Distributed hash table with the *n*-gram data
- Phase 2 calculate the cohesion
- Phase 3 identifies the *n*-grams that can be considered RE
- Ensures the same precision and recall of the LocalMaxs method definition
- An *n*-gram cache system, to reduce the remote data communication
- Analytical model to understand cache miss ratio and miss penalty
- *n*-gram repetition depends on:
  - Corpus size
  - Language
  - *n*-gram size

Results

- Extraction of relevant *2*-grams and *3*-grams exhibits almost linear speedup and sizeup
- The approach is scalable to larger *corpora* sizes and higher size *n*-grams by simply increasing the number of machines
- Cache usage can reduced the remote data communication, leading to 70% reduction in phase 2, and 55% reduction in the total execution time
- For each *corpus* size the number of distinct *n*-grams imposes a limit to the minimum remote communication overhead

ISEL NOVALINCS LABORATORY FOR COMPUTER SCIENCE AND INFORMATICS FCT FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE NOVA DE LISBOA FCT Fundação para a Ciência e a Tecnologia MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

Acknowledgments: FCT MCTES, NOVA LINCS UID/CEC/04516/2013; Lunacloud (http://www.lunacloud.com); ISEL "AzoresCloud"