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

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

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

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