Concurrent Processing of Text Corpus Queries

Radoslav Rábar, Pavel Rychlí

RASLAN 2015-12-04

This work has been partly supported by the Czech Science Foundation under the project GA15-13277S.
Motivation

- text corpora are huge collection of texts – up to billions of words
- queries evaluation can be slow
Manatee

- corpus manager
- implemented in C++
- FastStream and RangeStream
Go (golang)

- new programming language
- build-in concurrency primitives:
  - goroutine
  - channel
Implementation

- written in Go
- FastChan and RangeChan
- uses goroutines and channels
Figure: More cores caused worse performance.
Implementation – data exchange solution

Figure: Sending positions as batches improves performance.
Performance evaluation

- simple benchmark
- complex and quite extreme corpus queries
Performance evaluation

Figure: Compared time of the queries evaluation between the original and new implementation.
Performance evaluation - scalability of the new system

Figure: Scalability of the new implementation.

Speed up

<table>
<thead>
<tr>
<th>Cores</th>
<th>Speed up</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 core</td>
<td>+76 %</td>
</tr>
<tr>
<td>2 cores</td>
<td>+32 %</td>
</tr>
<tr>
<td>3 cores</td>
<td>+23 %</td>
</tr>
<tr>
<td>4 cores</td>
<td>+12 %</td>
</tr>
<tr>
<td>5 cores</td>
<td>+5 %</td>
</tr>
<tr>
<td>6 cores</td>
<td>+3 %</td>
</tr>
<tr>
<td>7 cores</td>
<td>-6 %</td>
</tr>
<tr>
<td>8 cores</td>
<td></td>
</tr>
</tbody>
</table>

Radoslav Rábara, Pavel Rychlí
Concurrent Processing of Text Corpus Queries
Comparing lines of code

- the original implementation
  - C++
  - iterators

- the new implementation
  - Go
  - goroutines and channels
  - does not implement all the functionality
Comparing lines of code

Figure: Graph comparing lines of code of the original and new implementation

Radoslav Rába, Pavel Rychly
Concurrent Processing of Text Corpus Queries
The new system has:

- better performance
- shorter source code

Therefore, it will replace the original system.
The new system has:
- better performance
- shorter source code

Therefore, it will replace the original system.