Hasso Plattner Institut

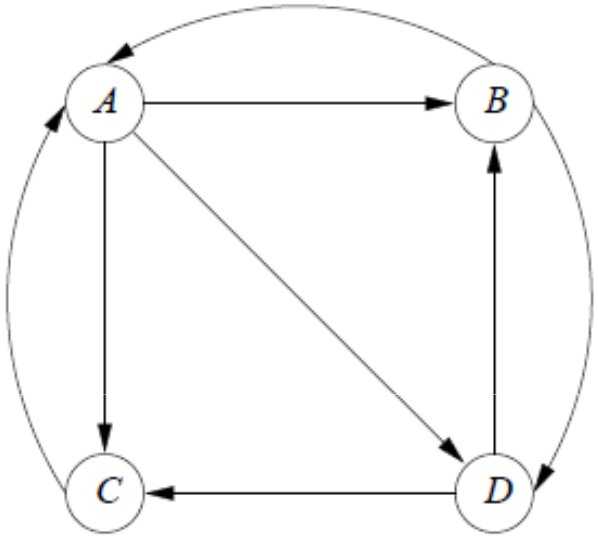
IT Systems Engineering | Universität Potsdam

Link Analysis goes MapReduce

The Hadoop Implementation

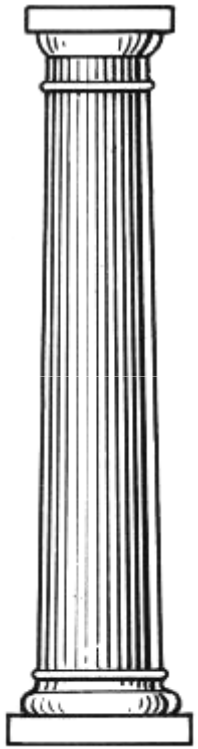
Thomas Zimmermann
Philipp Berger

Flashback

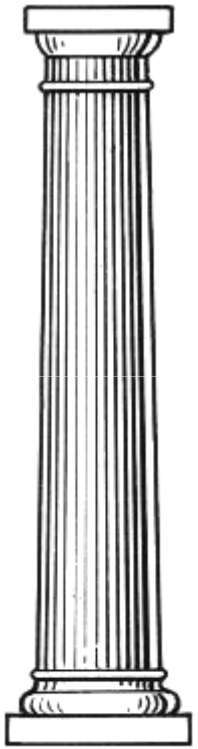


$$\begin{pmatrix} m_{1,1} & m_{1,2} & m_{1,3} & m_{1,4} \\ m_{2,1} & m_{2,2} & m_{2,3} & m_{2,4} \\ m_{3,1} & m_{3,2} & m_{3,3} & m_{3,4} \\ m_{4,1} & m_{4,2} & m_{4,3} & m_{4,4} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix}$$

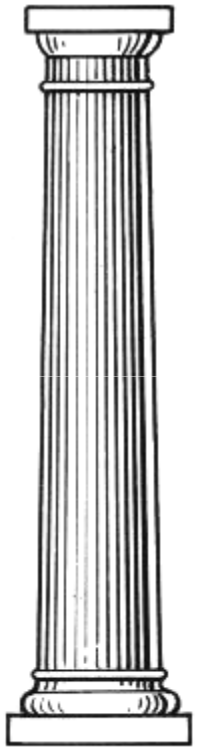
Overview



1. Pre- / Postprocessing

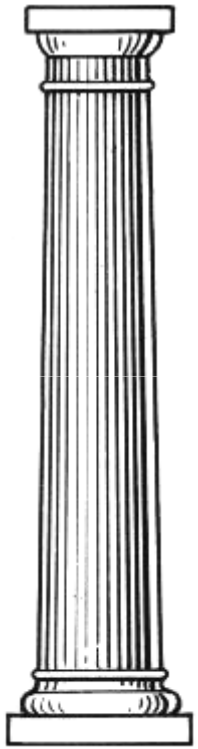


2. Our Jobs

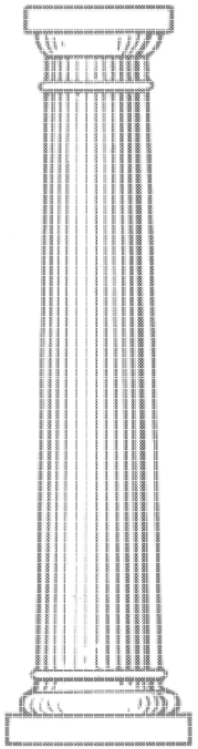


3. Evaluation

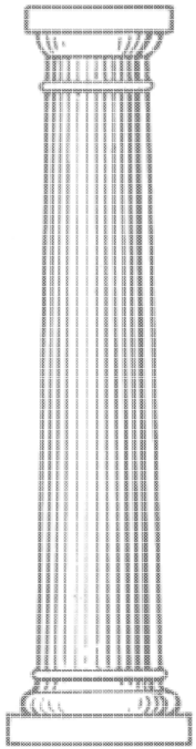
Overview



1. Pre- / Postprocessing



2. Our Jobs



3. Evaluation

Our Dataset

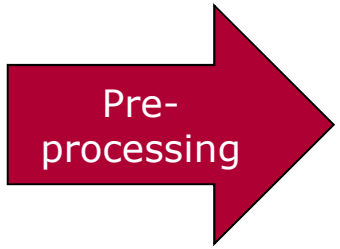
- DBPedia Infobox Properties

<http://dbpedia.org/resource/Eiffel_Tower>
<http://dbpedia.org/ontology/location>
<http://dbpedia.org/resource/France> .

<http://dbpedia.org/resource/Eiffel_Tower>
<http://dbpedia.org/ontology/location>
<http://dbpedia.org/resource/Paris> .

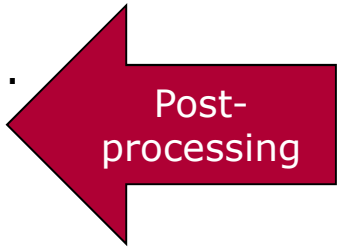
<http://dbpedia.org/resource/Eiffel_Tower>
<http://dbpedia.org/ontology/type>
<http://dbpedia.org/resource/Radio_broadcasting> .

<http://dbpedia.org/resource/Eiffel_Tower>
http://www.georss.org/georss/point
"48.8583 2.2945"@en .



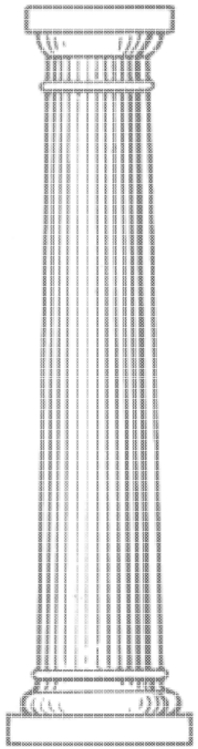
1, 2

1, 3

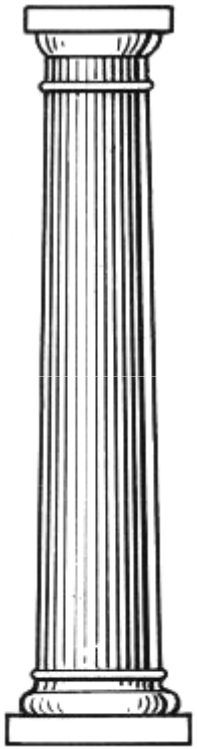


1, 4

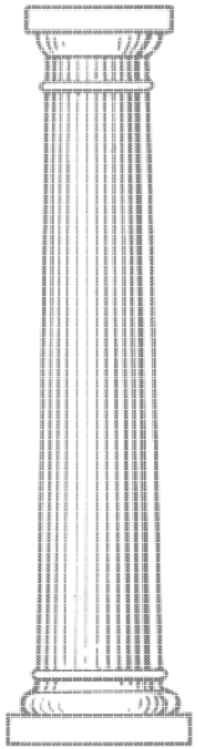
Overview



1. Pre- / Postprocessing

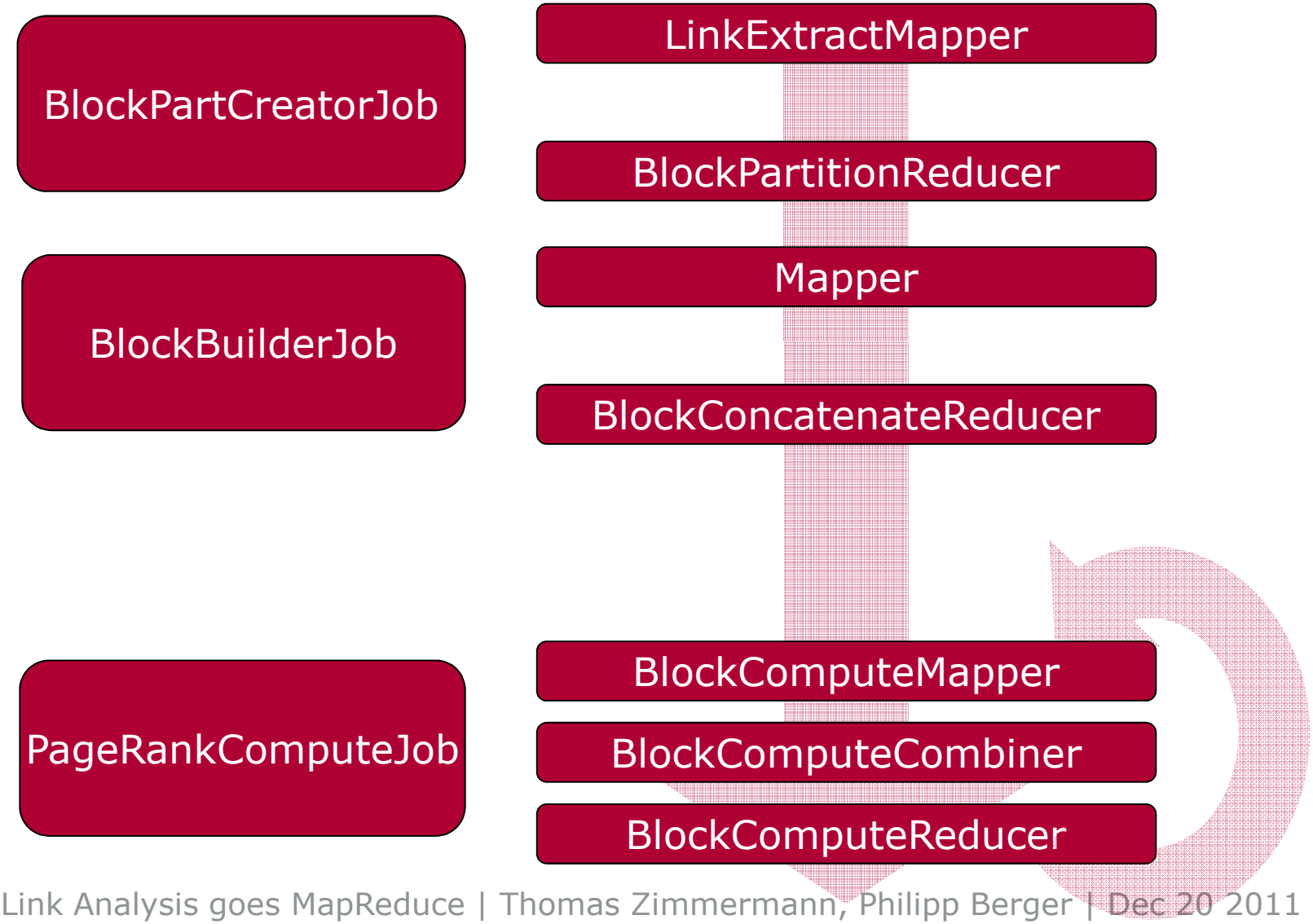


2. Our Jobs

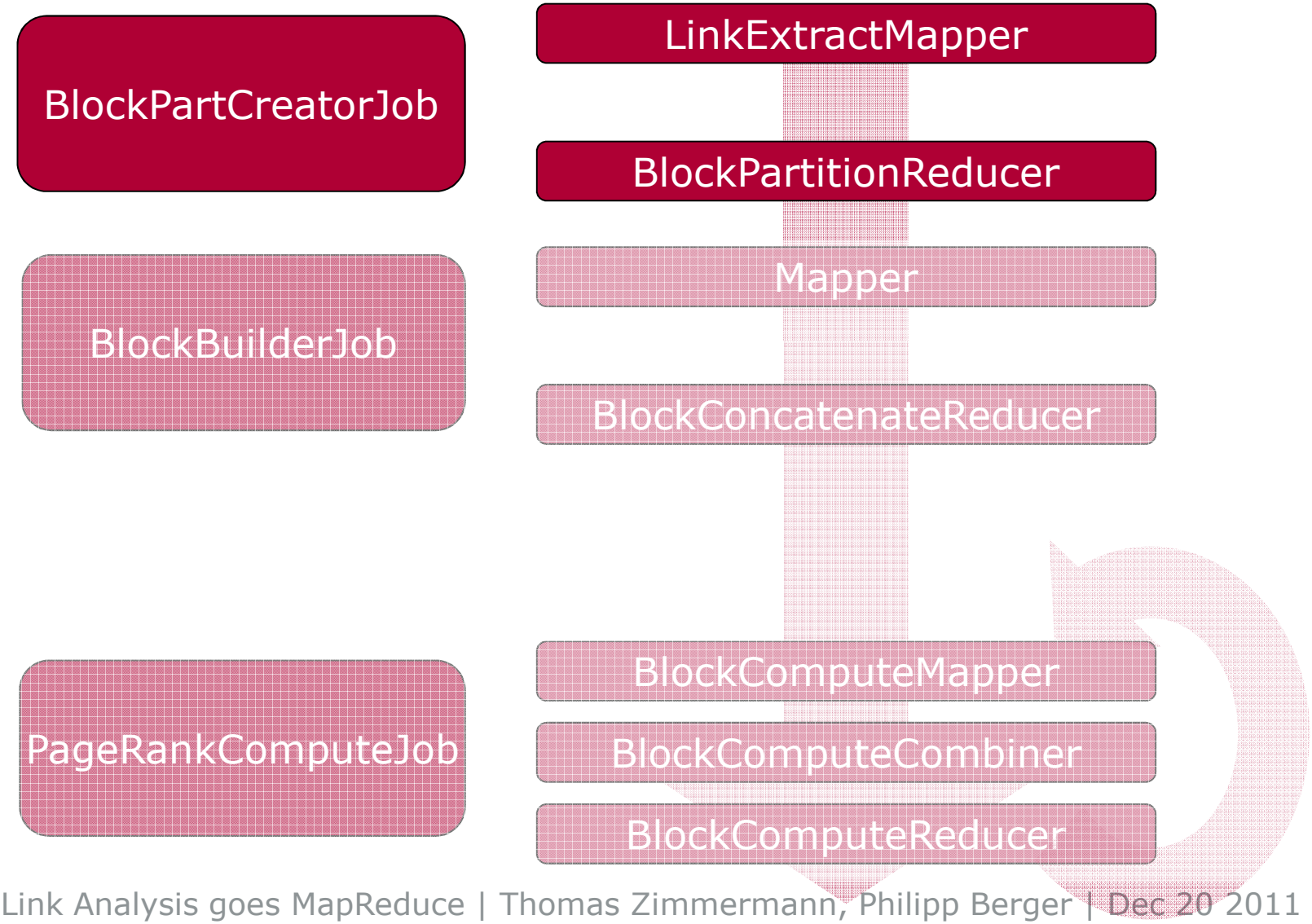


3. Evaluation

Our Jobs


BlockPartCreatorJob
BlockBuilderJob
PageRankComputeJob
LinkExtractMapper
BlockPartitionReducer
Mapper
BlockConcatenateReducer
BlockComputeMapper
BlockComputeCombiner
BlockComputeReducer

Our Jobs


BlockPartCreatorJob
LinkExtractMapper
BlockPartitionReducer
BlockBuilderJob
Mapper
BlockConcatenateReducer
PageRankComputeJob
BlockComputeMapper
BlockComputeCombiner
BlockComputeReducer

Our Jobs – Create Block Parts (Map)

LinkExtractMapper

Input:

Key, ($32_{(\text{Eiffel Tower})}$, $122_{(\text{France})}$)

Output:

$32_{(\text{Eiffel Tower})}$, $122_{(\text{France})}$

Our Jobs – Create Block Parts (Reduce)

BlockPartitionReducer

	$32_{\text{Eiffel Tower}}$	33_{Sushi}
122_{France}	X	
123_{Japan}		X
124_{Paris}	X	

Input:

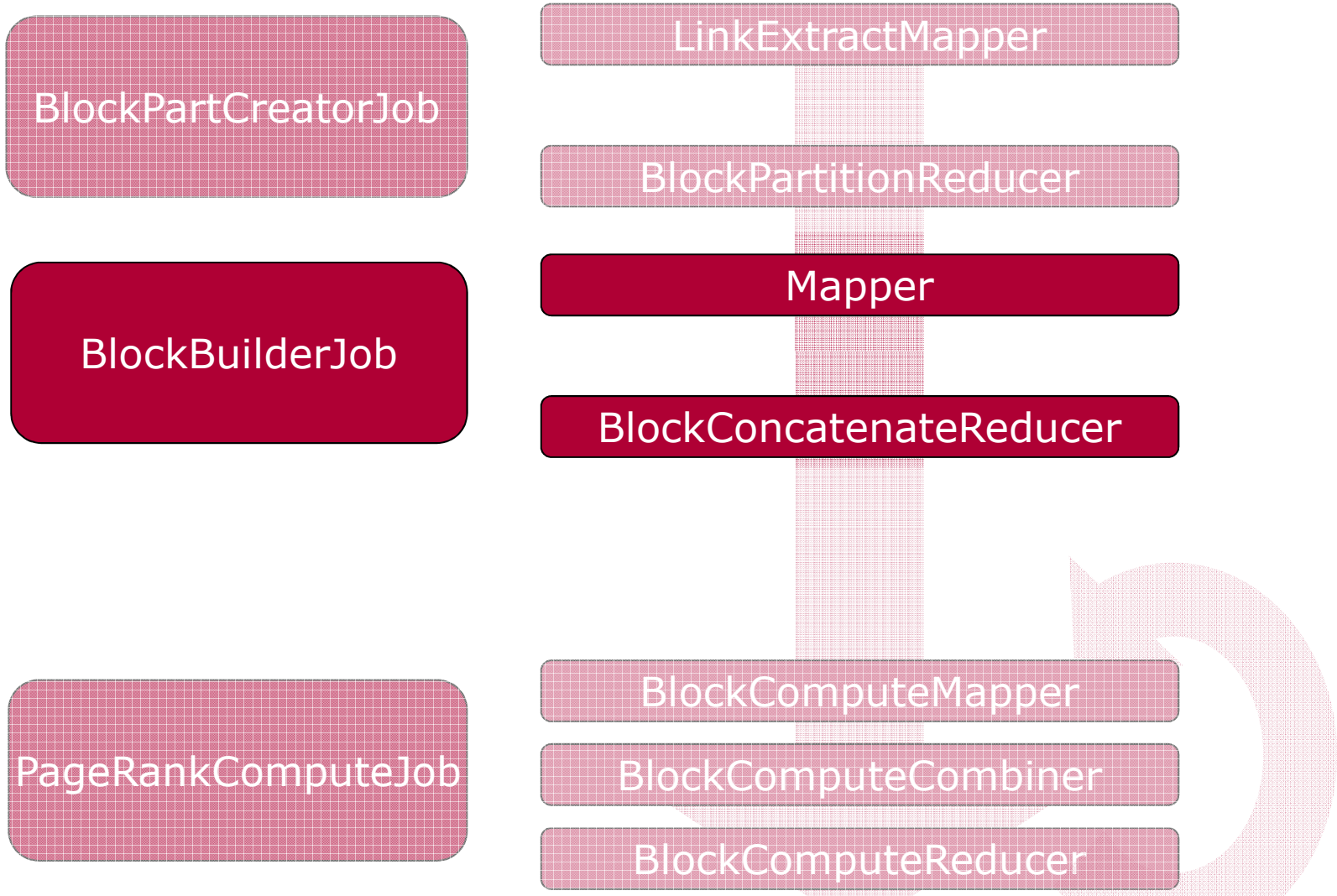
$32_{(\text{Eiffel Tower})}$, [$122_{(\text{France})}$, $124_{(\text{Paris})}$]

Output:

(16, 61) , ($32_{\text{Eiffel Tower}}$, 122_{France}, 2 #OutgoingLinks)

$\frac{32_{Eiffel\,Tower}}{2_{BlockSize}}$ $\frac{122_{France}}{2_{BlockSize}}$

Our Jobs


BlockPartCreatorJob
BlockBuilderJob
PageRankComputeJob
LinkExtractMapper
BlockPartitionReducer
Mapper
BlockConcatenateReducer
BlockComputeMapper
BlockComputeCombiner
BlockComputeReducer

Our Jobs – Concatenate Block Parts

BlockConcatenateReducer

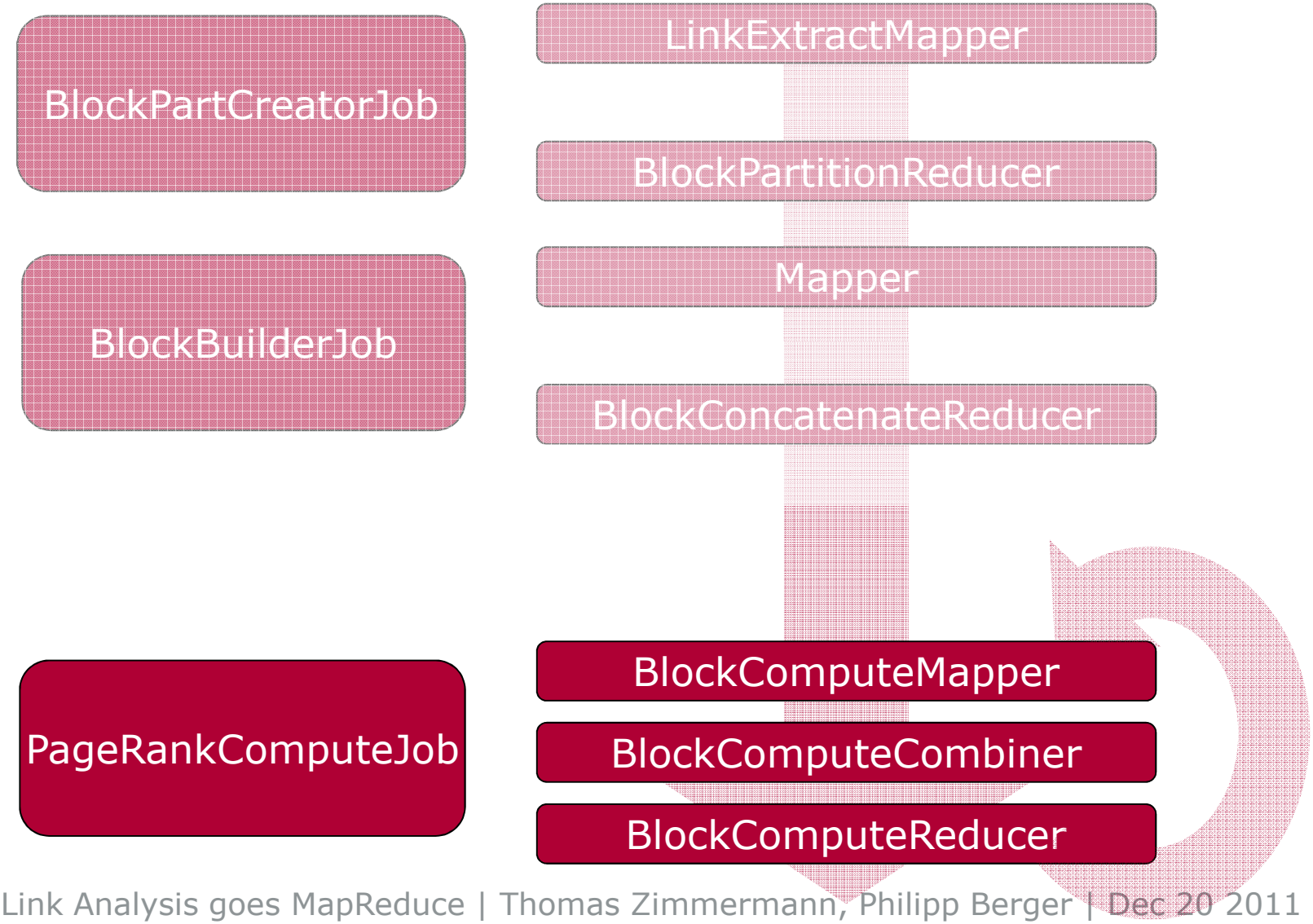
	$32_{\text{Eiffel Tower}}$	33_{Sushi}
122_{France}	X	
123_{Japan}		X
124_{Paris}	X	

Input / Output:

$(16_{\text{Eiffel Tower,Sushi}}, 61_{\text{France,Japan}})$,

$[(32_{\text{Eiffel Tower}}, 122_{\text{France}}, 2_{\#\text{OutgoingLinks}}), (33_{\text{Sushi}}, 123_{\text{Japan}}, 1_{\#\text{OutgoingLinks}})]$

Our Jobs


BlockPartCreatorJob
BlockBuilderJob
PageRankComputeJob
LinkExtractMapper
BlockPartitionReducer
Mapper
BlockConcatenateReducer
BlockComputeMapper
BlockComputeCombiner
BlockComputeReducer

Our Jobs – Do PageRank

BlockComputeMapper

Input:

$(16_{\text{Eiffel Tower,Sushi}}, 61_{\text{France,Japan}})$,

$((32_{\text{Eiffel Tower}}, 122_{\text{France}}, 2_{\#\text{OutgoingLinks}}), (33_{\text{Sushi}}, 123_{\text{Japan}}, 1_{\#\text{OutgoingLinks}}))$

- Get PageRank of sources for one Block

Output:

122_{France}, rank(Eiffel Tower) * 1/2

123_{Japan}, rank(Sushi) * 1/1

Distributed Cache

Page	Rank
$32_{\text{Eiffel Tower}}$	0.125
33_{Sushi}	0.125
34_{Island}	0.125

Our Jobs – Do PageRank

BlockComputeCombiner

Input:

122_{France}, rank(Eiffel Tower) * ½

122_{France}, rank(Paris)* ¼

- Simply sums up incoming values

Output:

122_{France}, rank(Eiffel Tower) * ½ + rank(Paris) * ¼ }

Our Jobs – Do PageRank

BlockComputeReducer

Input:

122_{France}, rank(Eiffel Tower) * ½ + rank(Paris) * ¼

122_{France}, rank(Lyon) * ¼

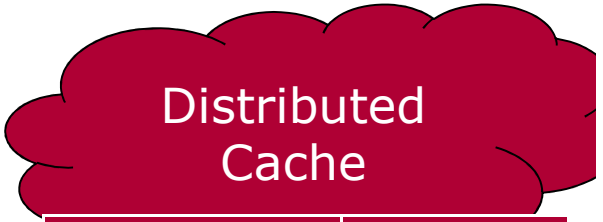
```
// Random Jump
result = beta * pageRankSum + (1 - beta) * (1 / pageCount);
```

Output:

122_{France}, result

Our Jobs – Do PageRank

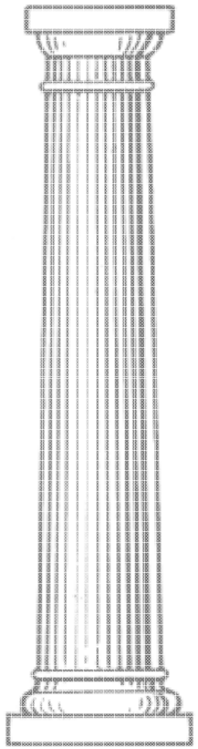
- Distributed Cache Update
  - Read the output of PageRank calculation
  - Load the PageRank into the Distributed Cache



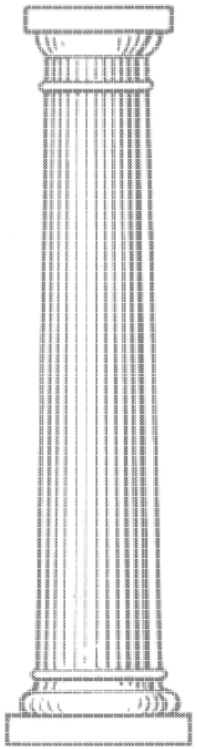
Page	Rank
$32_{\text{Eiffel Tower}}$	0.105
33_{Sushi}	0.025
34_{Island}	0.177

- Start next iteration with new PageRank in Distributed Cache
  - Until **convergence** is reached
    - Until the root-mean-square-deviation < target variance

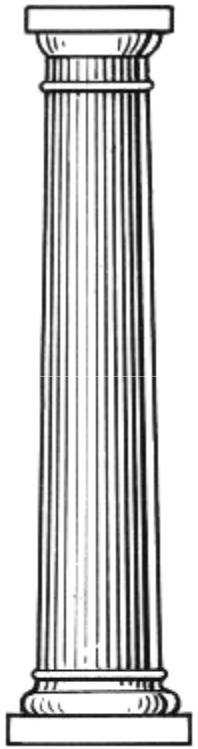
Overview



1. Pre- / Postprocessing



2. Our Jobs



3. Evaluation

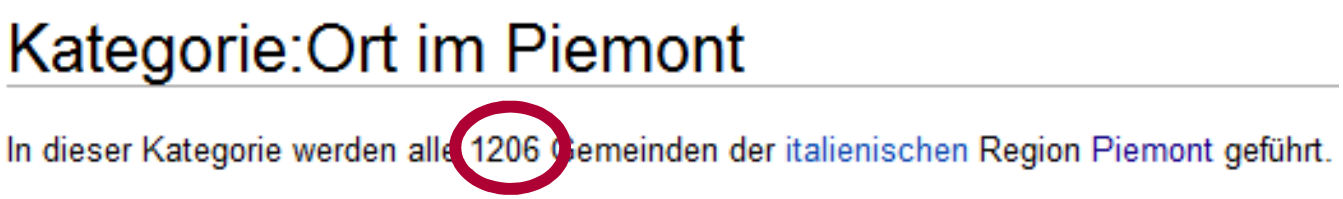
Results: Top 10 German Sites

Rank	PageRank	Site
1	0.0084	<http://www.opengis.net/gml/_Feature>
2	2.9127E-4	<http://dbpedia.org/resource/Rheinland-Pfalz>
3	2.4293E-4	<http://dbpedia.org/resource/Deutschland>
4	2.3606E-4	<http://dbpedia.org/resource/Vereinigte_Staaten>
5	1.4299E-4	<http://dbpedia.org/resource/Piemont>
6	1.4109E-4	<http://dbpedia.org/resource/Schleswig-Holstein>
7	1.3978E-4	<http://dbpedia.org/resource/Texas>
8	1.3291E-4	<http://dbpedia.org/resource/United_States>
9	1.3159E-4	<http://dbpedia.org/resource/Baden-Württemberg>
10	1.2901E-4	<http://dbpedia.org/resource/Niedersachsen>

Results: Top 10 German Sites

- 2477 pages link to „Rheinland-Pfalz“!

- 1242 to „Piemont“ (Italy‘s largest region)



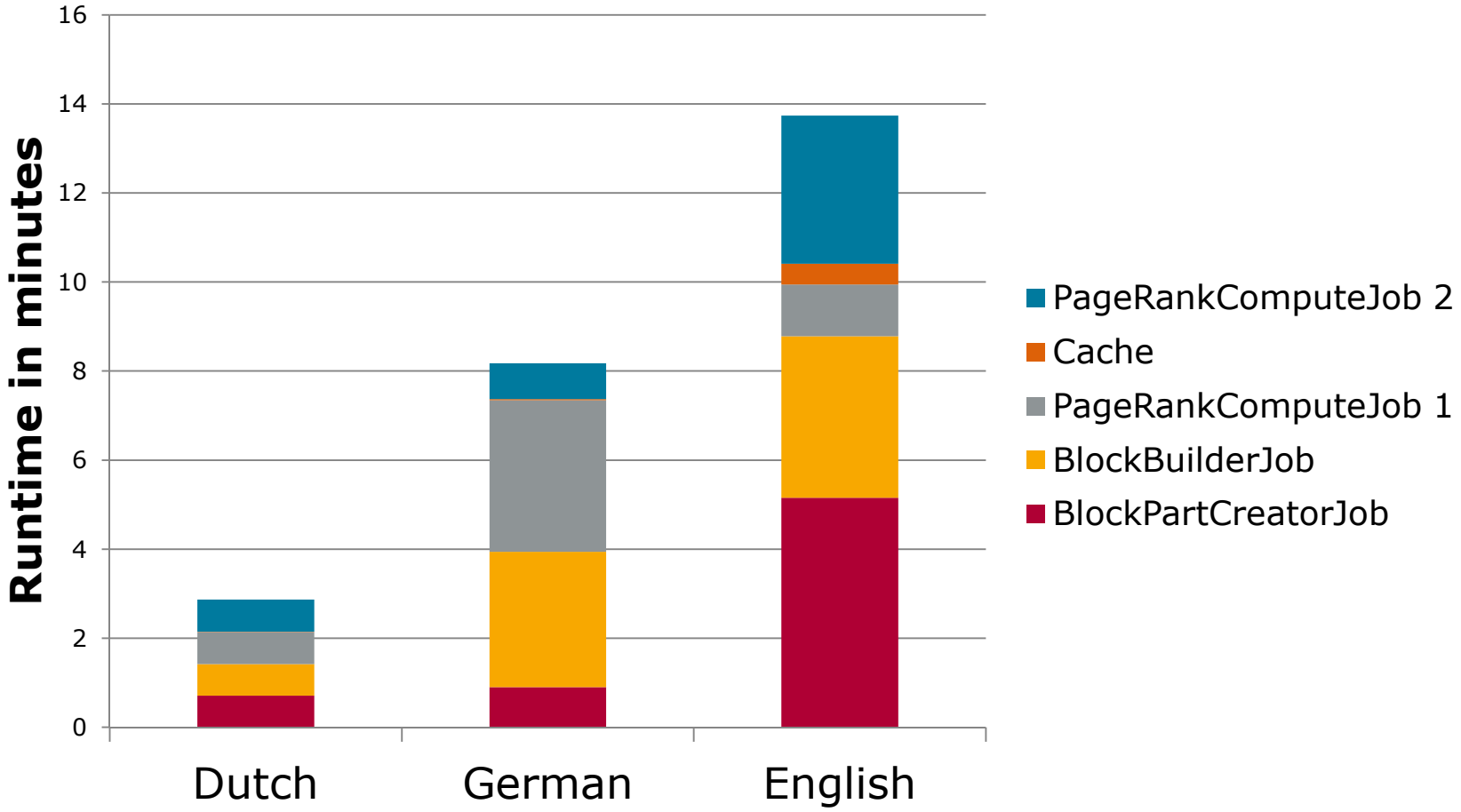
Rank	Site
1	<http://www.opengis.net/gml/_Feature>
2	<http://dbpedia.org/resource/Rheinland-Pfalz>
3	<http://dbpedia.org/resource/Deutschland>
4	<http://dbpedia.org/resource/Vereinigte_Staaten>
5	<http://dbpedia.org/resource/Piemont>
6	<http://dbpedia.org/resource/Schleswig-Holstein>
7	<http://dbpedia.org/resource/Texas>
8	<http://dbpedia.org/resource/United_States>
9	<http://dbpedia.org/resource/Baden-Württemberg>
10	<http://dbpedia.org/resource/Niedersachsen>

Evaluation

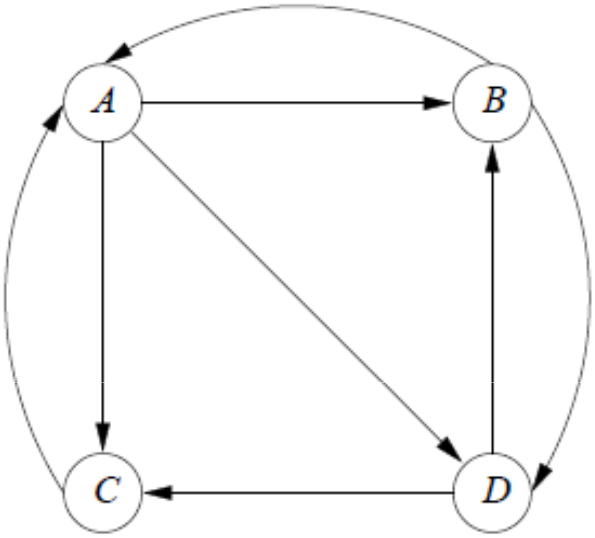
Data	SingleThreaded	Cluster
Dutch 1.9 MB	2 sec	2.53min
German 154.8 MB	4 min	8.25min
English 2.2 GB	>3 Std	14.22min

- Single-Threaded graph analysis library
  - edu.uci.ics.jung.algorithms.scoring.PageRank
- On Cluster with 10 nodes
  - Blocksize 10 000

Evaluation


Runtime in minutes
16
14
12
10
8
6
4
2
0
Dutch
German
English
PageRankComputeJob 2
Cache
PageRankComputeJob 1
BlockBuilderJob
BlockPartCreatorJob

Summary



$$\begin{pmatrix} m_{1,1} & m_{1,2} & m_{1,3} & m_{1,4} \\ m_{2,1} & m_{2,2} & m_{2,3} & m_{2,4} \\ m_{3,1} & m_{3,2} & m_{3,3} & m_{3,4} \\ m_{4,1} & m_{4,2} & m_{4,3} & m_{4,4} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix}$$

BlockPartCreatorJob

BlockBuilderJob

PageRankComputeJob