Similarity Problems in Paragraph Justification: an Extension to the Knuth-Plass Algorithm

TUG 2024

Didier Verna
EPITA / LRE

didier@lrde.epita.fr

lrde/~didier


@didierverna


didier.verna


in/didierverna

Conclusion

- An extension to the Knuth-Plass algorithm
- Addresses similarity problems in paragraph justification

Thank you!

Similarities

In olden times when wishing still helped one, there lived a king whose daughters were all beautiful; and the youngest was so beautiful that the sun itself, which has seen so much, was astonished whenever it shone in her face. Close by the king's castle lay a great dark forest, and under an old lime-tree in the forest was a well, and when the day was very warm, the king's child went out into the forest and sat down by the side of the cool fountain; and when she was bored she took a golden ball, and threw it up on high and caught it; and this ball was her favorite plaything.

- TEX's rendition
  - `\linewidth=201pt`
  - `\font \tenlmr=ec-lmr10`

Similarities

In olden times when wishing still helped one, there lived a king whose daughters were all beautiful; and the youngest was so beautiful that the sun itself, which has seen so much, was astonished whenever it shone in her face. Close by the king's castle lay a great dark forest, and under an old lime-tree in the forest was a well, and when the day was very warm, the king's child went out into the forest and sat down by the side of the cool fountain; and when she was bored she took a golden ball, and threw it up on high and caught it; and this ball was her favorite plaything.

- TEX's rendition
  - `\linewidth=201pt`
  - `\font \tenlmr=ec-lmr10`
- Problems (also at beginning of line)
  - Micro-interruption
  - Accidental line skipping

Similarities

In olden times when wishing still helped one, there lived a king whose daughters were all beautiful; and the youngest was so beautiful that the sun itself, which has seen so much, was astonished whenever it shone in her face. Close by the king's castle lay a great dark forest, and under an old lime-tree in the forest was a well, and when the day was very warm, the king's child went out into the forest and sat down by the side of the cool fountain; and when she was bored she took a golden ball, and threw it up on high and caught it; and this ball was her favorite plaything.

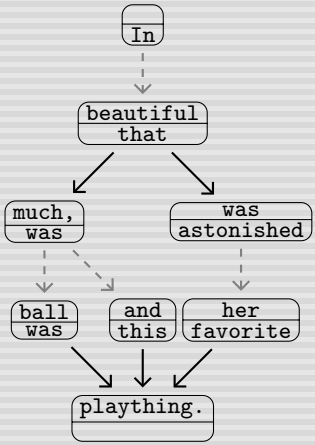
- TEX's rendition
  - `\linewidth=201pt`
  - `\font \tenlmr=ec-lmr10`
- Problems (also at beginning of line)
  - Micro-interruption
  - Accidental line skipping
- Ancient: scribal errors (monks)

Similarities

In olden times when wishing still helped one, there lived a king whose daughters were all beautiful; and the youngest was so beautiful that the sun itself, which has seen so much, was astonished whenever it shone in her face. Close by the king's castle lay a great dark forest, and under an old lime-tree in the forest was a well, and when the day was very warm, the king's child went out into the forest and sat down by the side of the cool fountain; and when she was bored she took a golden ball, and threw it up on high and caught it; and this ball was her favorite plaything.

- TEX's rendition
  - `\linewidth=201pt`
  - `\font \tenlmr=ec-lmr10`
- Problems (also at beginning of line)
  - Micro-interruption
  - Accidental line skipping
- Ancient: scribal errors (monks)
- Terminology
  - "Saut du même au même" (french)
  - Homeoarchy / homeoteleuton (rhetoric)
  - Character / word ladders ?

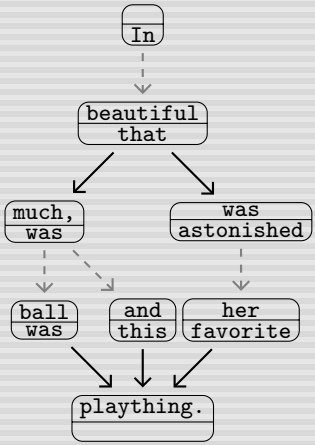
The Knuth-Plass in a Nutshell



- "Single Pair Shortest Path" problem

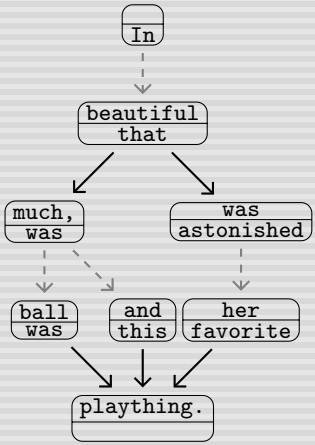
  Finding the "best" route from beginning to end of paragraph

The Knuth-Plass in a Nutshell



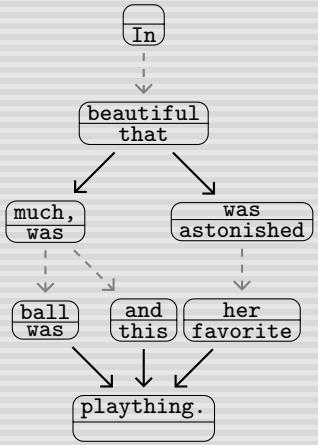
- "Single Pair Shortest Path" problem
  Finding the "best" route from beginning to end of paragraph
- Dynamic optimization technique
  Never construct the full graph entirely

The Knuth-Plass in a Nutshell



- "Single Pair Shortest Path" problem
  Finding the "best" route from beginning to end of paragraph
- Dynamic optimization technique
  Never construct the full graph entirely
- Cost function
  - Local demerits: line by line
    Badness, hyphenation penalty, etc.
  - Contextual demerits: line against line
    Hyphenation ladders, stretching / shrinking disparities, etc.

The Knuth-Plass in a Nutshell



- "Single Pair Shortest Path" problem
  Finding the "best" route from beginning to end of paragraph
- Dynamic optimization technique
  Never construct the full graph entirely
- Cost function
  - Local demerits: line by line
    Badness, hyphenation penalty, etc.
  - Contextual demerits: line against line
    Hyphenation ladders, stretching / shrinking disparities, etc.
- Decision: $\min \sum(\textit{local} + \textit{contextual demerits})$

Knuth-Plass Extension

- **Idea:** a new kind of contextual demerits (“similar demerits”)
  - Compare the beginning & end of consecutive lines
  - Add similar demerits to the total if identical

Knuth-Plass Extension

- **Idea:** a new kind of contextual demerits (“similar demerits”)
  - Compare the beginning & end of consecutive lines
  - Add similar demerits to the total if identical
  - Note: special case for the final 2 lines

Knuth-Plass Extension

- **Idea:** a new kind of contextual demerits (“similar demerits”)
  - Compare the beginning & end of consecutive lines
  - Add similar demerits to the total if identical
  - Note: special case for the final 2 lines
- **Implementation:** nodes include the bol / eol characters
  - discarding kerns
  - up to the first glue or discretionary

Knuth-Plass Extension

- **Idea:** a new kind of contextual demerits (“similar demerits”)
  - Compare the beginning & end of consecutive lines
  - Add similar demerits to the total if identical
  - Note: special case for the final 2 lines
- **Implementation:** nodes include the bol / eol characters
  - discarding kerns
  - up to the first glue or discretionary
- **Rationale:**
  - Kerns: small adjustments, identical in similarities
  - Discretionaries: avoid deconstruction overhead
  - Glue: avoid vertical alignment considerations

Knuth-Plass Extension

- **Idea:** a new kind of contextual demerits (“similar demerits”)
  - Compare the beginning & end of consecutive lines
  - Add similar demerits to the total if identical
  - Note: special case for the final 2 lines
- **Implementation:** nodes include the bol / eol characters
  - disc
  - up t
- **Rationale:**
  - Kerns: small adjustments, identical in similarities
  - Discretionaries: avoid deconstruction overhead
  - Glue: avoid vertical alignment considerations

Comparing two short sequences of characters

He who can do less can do more…

Demo

- Applied for $\mathit{length}(\mathit{similarity}) \geq 2$, calibration: `[0,10000]`

Demo

- Applied for $\mathit{length}(\mathit{similarity}) \geq 2$, calibration: `[0,10000]`

In olden times when wishing still helped one, there lived a king whose daughters were all beautiful; and the youngest was so beautiful that the sun itself, which has seen so much, was astonished whenever it shone in her face. Close by the king's castle lay a great dark forest, and under an old lime-tree in the forest was a well, and when the day was very warm, the king's child went out into the forest and sat down by the side of the cool fountain; and when she was bored she took a golden ball, and threw it up on high and caught it; and this ball was her favorite plaything.

`\similardemerits0`

In olden times when wishing still helped one, there lived a king whose daughters were all beautiful; and the youngest was so beautiful that the sun itself, which has seen so much, was astonished whenever it shone in her face. Close by the king's castle lay a great dark forest, and under an old lime-tree in the forest was a well, and when the day was very warm, the king's child went out into the forest and sat down by the side of the cool fountain; and when she was bored she took a golden ball, and threw it up on high and caught it; and this ball was her favorite plaything.

`\similardemerits2800`

In olden times when wishing still helped one, there lived a king whose daughters were all beautiful; and the youngest was so beautiful that the sun itself, which has seen so much, was astonished whenever it shone in her face. Close by the king's castle lay a great dark forest, and under an old lime-tree in the forest was a well, and when the day was very warm, the king's child went out into the forest and sat down by the side of the cool fountain; and when she was bored she took a golden ball, and threw it up on high and caught it; and this ball was her favorite plaything.

`\similardemerits5230`

Is it worth it?

- **Pertinence:** is the problem frequent?
- **Efficacy:** does this Knuth-Plass extension solve the problem?

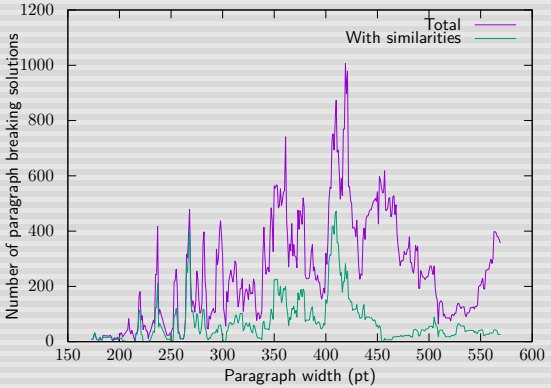
Is it worth it?

- **Pertinence:** is the problem frequent?
- **Efficacy:** does this Knuth-Plass extension solve the problem?
- **Two typesetting experiments:**
  - One paragraph at many different widths
  - Many paragraphs at a single width

Pertinence

Experimental conditions

- Grimm Brothers “Frog King” §1
- 142pt (≈ 5cm) → 569pt (≈ 20cm)
- 427 runs



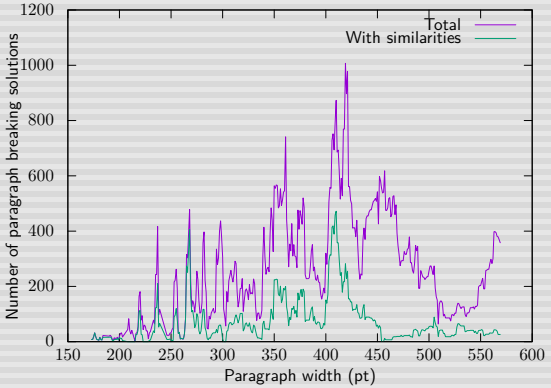
Pertinence

Experimental conditions

- Grimm Brothers “Frog King” §1
- 142pt (≈ 5cm) → 569pt (≈ 20cm)
- 427 runs

Expected results

- Chaotic shapes
- Close numbers for narrow paragraphs



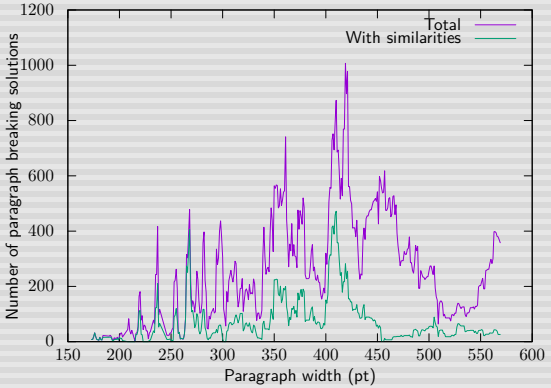
Pertinence

Experimental conditions

- Grimm Brothers “Frog King” §1
- 142pt (≈ 5cm) → 569pt (≈ 20cm)
- 427 runs

Findings

- Similarities are frequent
- So are similarity-free layouts



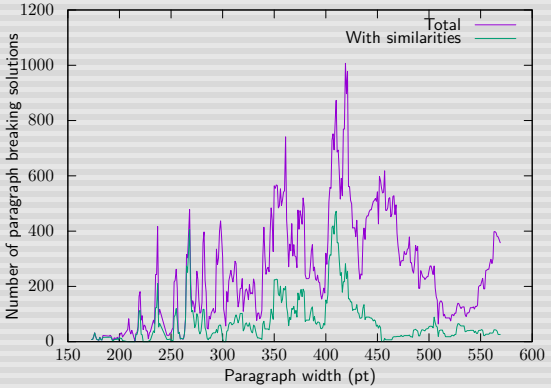
Pertinence

Experimental conditions

- Grimm Brothers “Frog King” §1
- 142pt (≈ 5cm) → 569pt (≈ 20cm)
- 427 runs

What about TEX?

- 4%: similarities unavoidable
- 21%: with similarities
- 26%: experiment 2



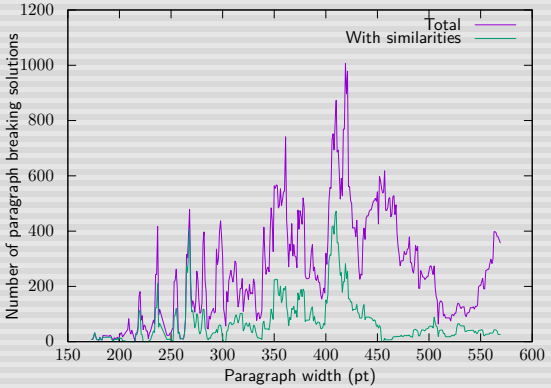
Pertinence

Experimental conditions

- Grimm Brothers “Frog King” §1
- 142pt (≈ 5cm) → 569pt (≈ 20cm)
- 427 runs

Conclusion

- 2 paragraphs out of 10 is a lot
- The problem is worth addressing



Efficacy

Experimental conditions

- Herman Melville's Mobby Dick
- 1524 paragraphs @ 284pt ($\approx$ 10cm)
- $+$ experiment 1 $\rightarrow$ 1951 cases
- $x3 \rightarrow$ 5853 runs

`\similardemerits10000`	$+$ `\adjacentdemerits0`
Corrected: 48% / 50% Improved: 50% / 63%	Corrected: 53% / 66% Improved: 57% / 73%

Efficacy

Experimental conditions

- Herman Melville's Mobby Dick
- 1524 paragraphs @ 284pt ($\approx$ 10cm)
- $+$ experiment 1 $\rightarrow$ 1951 cases
- $x3 \rightarrow$ 5853 runs

Conclusion

- Automatic treatment possible

`\similardemerits10000`	$+$ `\adjacentdemerits0`
Corrected: 48% / 50% Improved: 50% / 63%	Corrected: 53% / 66% Improved: 57% / 73%

Efficacy

Experimental conditions

- Herman Melville's Mobby Dick
- 1524 paragraphs @ 284pt ($\approx$ 10cm)
- $+$ experiment 1 $\rightarrow$ 1951 cases
- $x3 \rightarrow$ 5853 runs

Conclusion

- Automatic treatment possible

`\similardemerits10000`	$+$ `\adjacentdemerits0`
Corrected: 48% / 50% Improved: 50% / 63%	Corrected: 53% / 66% Improved: 57% / 73%

Please note

- Those are extreme conditions!
- Similarity-free $\neq$ aesthetically better
- But we have some leeway...

(Real) Conclusion

- Similarity handling for the Knuth-Plass
  - Simple & lightweight
  - Backward-compatible with regular TEX (`\similardemerits0`)

(Real) Conclusion

- Similarity handling for the Knuth-Plass
  - Simple & lightweight
  - Backward-compatible with regular TEX (`\similardemerits0`)
- Implemented in ETAP, could be useful in production engines

(Real) Conclusion

- Similarity handling for the Knuth-Plass
  - Simple & lightweight
  - Backward-compatible with regular TEX (`\similardemerits0`)
- Implemented in ETAP, could be useful in production engines
- Experimentation → automatic similarity handling is both worthy and achievable

(Real) Conclusion

- Similarity handling for the Knuth-Plass
  - Simple & lightweight
  - Backward-compatible with regular TEX (`\similardemerits0`)
- Implemented in ETAP, could be useful in production engines
- Experimentation → automatic similarity handling is both worthy and achievable

Perspectives

- Study trade-off with other aesthetic criteria
- Similar demerits as $f(\textit{similarity})$ rather than just a constant
- Currently talking with Hans... 😉