Line-breaking Algorithm in Tex and Its Future

7/27/24

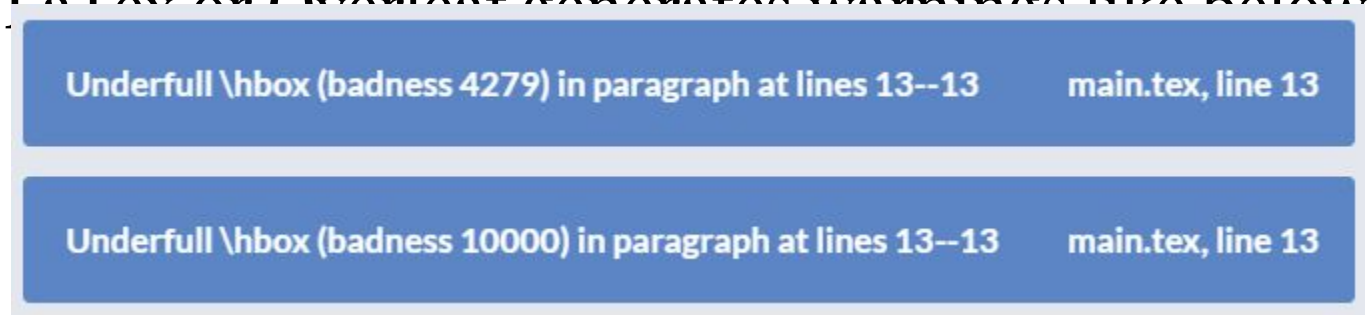
Qingyang Zhou
Cryptography, Security, and Privacy (CrySP)

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

Do you have the situation that...

1. The conference requires a 13 page-limit paper, and your paper is 13 + 1 line long? You delete some words just to fit the limitation, but the line does not remove at all?

2. LaTex or Overleaf generates warnings like below?



3. LaTex will generate longer lines even if you delete some words?

I will give you a very risky trick in the end. Use it at your own risk!

Outline

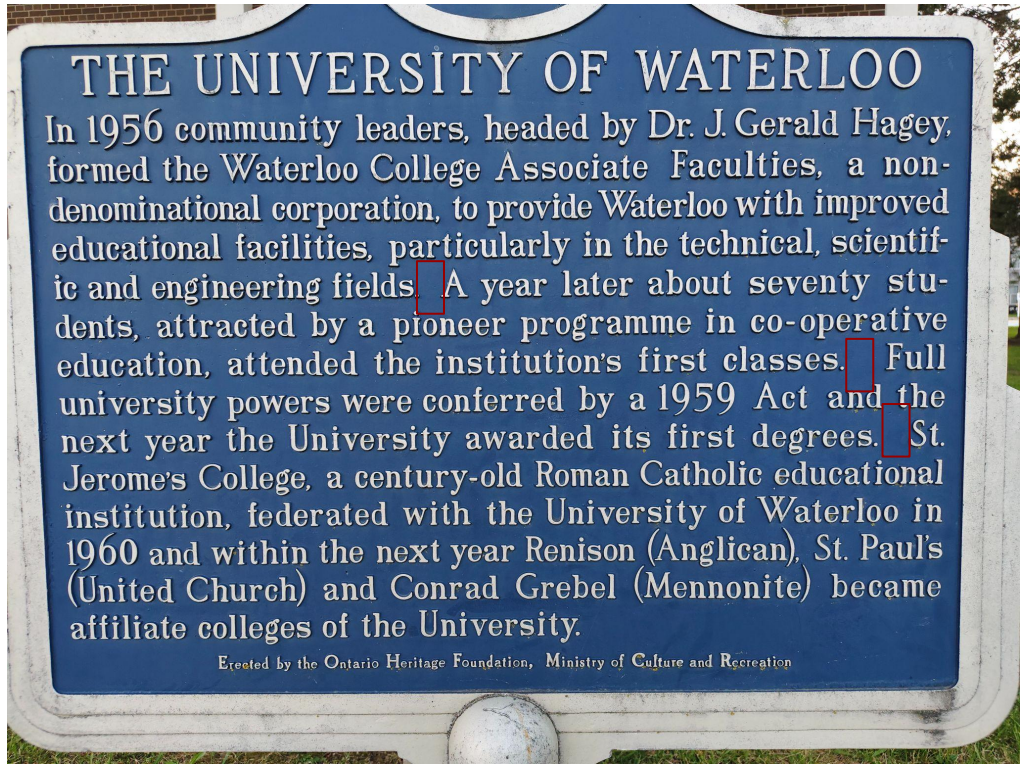
- Basic Knowledge about the line-breaking algorithm
- The line-breaking algorithm in Tex: Knuth–Plass Line-breaking Algorithm.
- The improvement of the K-P algorithm
- Future work

Basic Knowledge – Importance about Line-breaking

Breaking the paragraphs into different lines,
or the line-breaking problem, exists even before the beginning of electronic publishing.

Back in the day when people were still using typewriters, the typers needed to break the paragraphs manually by using bell rings.

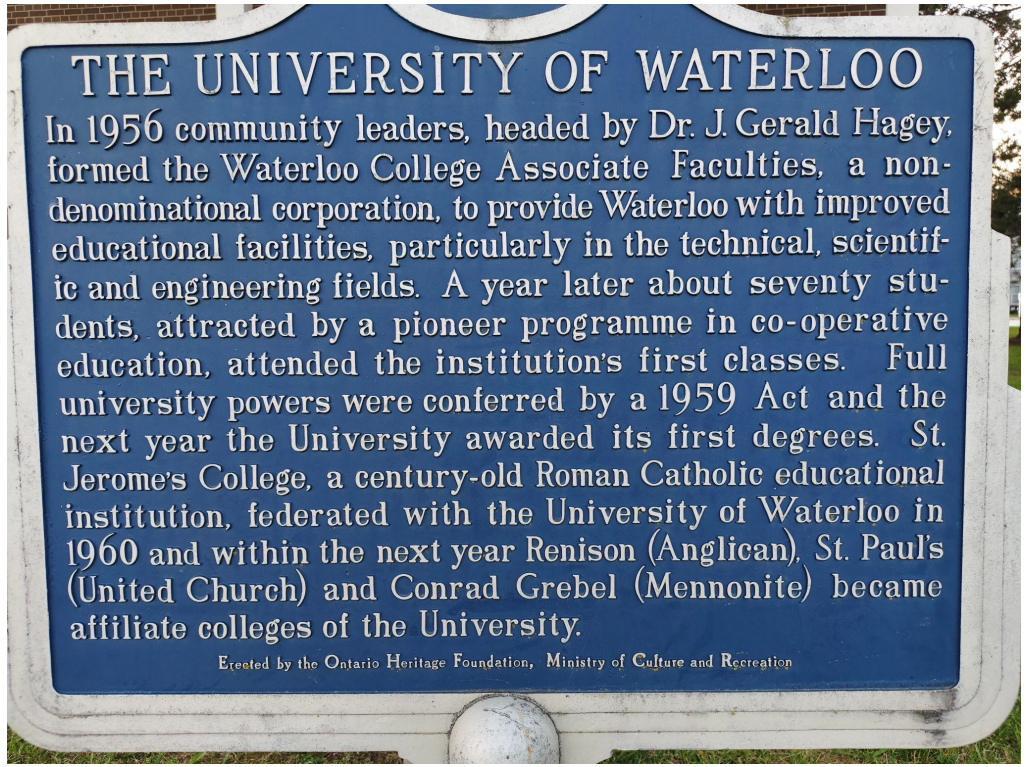
Bad line-breaking method distracts readers heavily. Look at the red rectangle right.



Basic Knowledge – Comparison

In 1956 community leaders, headed by Dr. J. Gerald Hagey, formed the Waterloo College Associate Faculties. a non-denominational corporation, to provide Waterloo with improved educational facilities, particularly in the technical, scientific and engineering fields. A year later about seventy students, attracted by a pointer programme in co-operative education, attended the institution's first classes. Full university powers were conferred by a 1959 Act and the next year the University awarded its first degrees. St. Jerome's College, a century-old Roman Catholic educational institution, federated with the University of Waterloo in 1960 and within the next year Renison (Anglican), St. Paul's (United Church) and Conrad Grebel (Mennonite) became affiliate colleges of the University.

latex



Boar
d

Basic Knowledge – What is Line-breaking?

But to solve the problem, we still need a *formal* definition of the paragraph and line.

A *Paragraph* is a sequence of x_1 … x_m items, where each x_i could be a *box*, a *glue* or a *penalty*.

A *Line* is a subsequence of a paragraph, which is x_i … x_j where 1 <= i <=j <= m. Each line has its ideal length l_i, which stands for the ideal horizontal space occupied by the line. Usually they are the same constant.

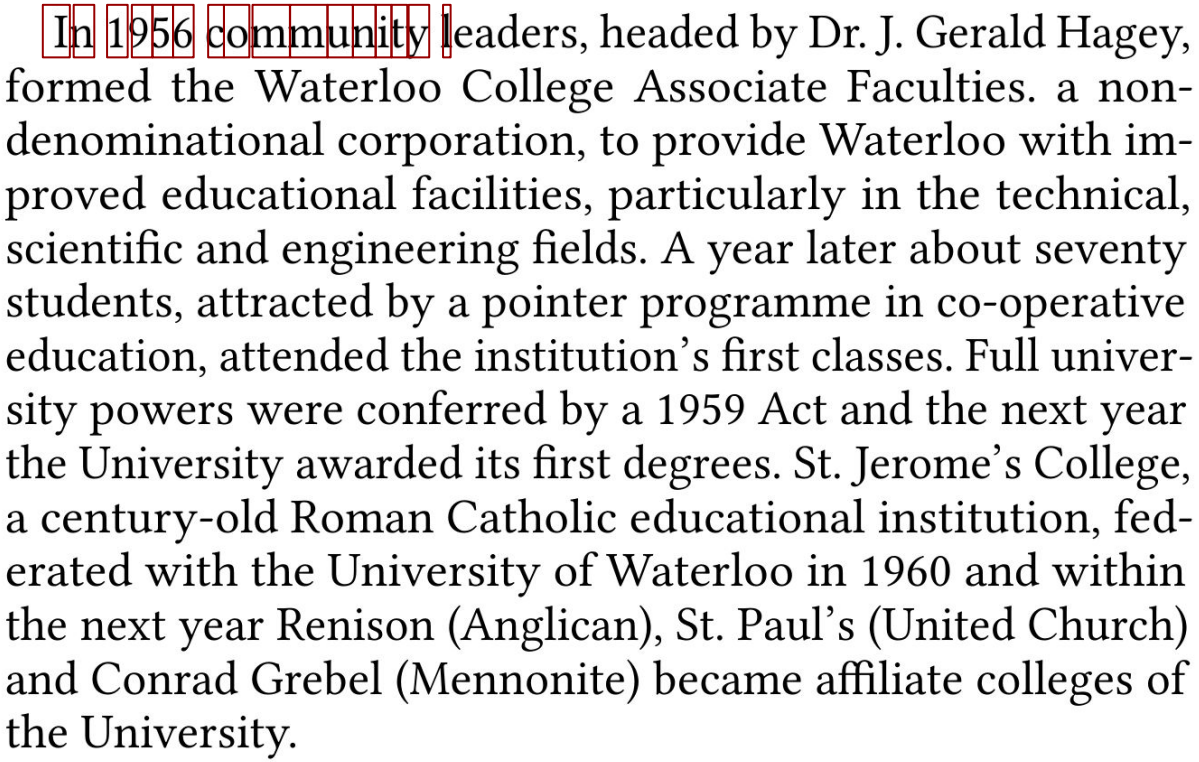
We will talk about box, glue and penalty in following.

Basic Knowledge – Box

Box refers to something that is to be typeset.

Box could be: a character from some font of type;
a black rectangle such as a horizontal or vertical rule;
as an accented letter;
a mathematical formula, etc.

the only relevant thing about a box is its *width*. Red in right is some box items.


In 1956 community leaders, headed by Dr. J. Gerald Hagey, formed the Waterloo College Associate Faculties. a non-denominational corporation, to provide Waterloo with improved educational facilities, particularly in the technical, scientific and engineering fields. A year later about seventy students, attracted by a pointer programme in co-operative education, attended the institution's first classes. Full university powers were conferred by a 1959 Act and the next year the University awarded its first degrees. St. Jerome's College, a century-old Roman Catholic educational institution, federated with the University of Waterloo in 1960 and within the next year Renison (Anglican), St. Paul's (United Church) and Conrad Grebel (Mennonite) became affiliate colleges of the University.

Basic Knowledge – Glue

Glue refers to blank space that can vary its width in specified ways.

The word ‘glue’ is perhaps not the best term, because it sounds a bit messy; a word like ‘spring’ would be better, since metal springs expand or compress to fill up space in essentially the way we want.

There are three real numbers (w, y, z) to describe glue: ‘normal’ width, ‘stretchability’ and ‘shrinkability’.

In 1956 community leaders, headed by Dr. J. Gerald Hagey, formed the Waterloo College Associate Faculties. a non-denominational corporation, to provide Waterloo with improved educational facilities, particularly in the technical, scientific and engineering fields. A year later about seventy students, attracted by a pointer programme in co-operative education, attended the institution’s first classes. Full university powers were conferred by a 1959 Act and the next year the University awarded its first degrees. St. Jerome’s College, a century-old Roman Catholic educational institution, federated with the University of Waterloo in 1960 and within the next year Renison (Anglican), St. Paul’s (United Church) and Conrad Grebel (Mennonite) became affiliate colleges of the University.

Basic Knowledge – Penalty

Penalty specifications refer to potential places to end one line of a paragraph and begin another.

There are three real numbers (p, w, f) to describe penalty.
p helps us decide whether or not to end a line at this point. infinity forbids break.
-infinity means a must break.
If a line break occurs at this place in the paragraph, additional width w will be added to the line.
Penalty specifications are of two kinds, flagged and unflagged, denoted by f.

In 1956 community leaders, headed by Dr. J. Gerald Hagey, formed the Waterloo College Associate Faculties. a non-denominational corporation, to provide Waterloo with improved educational facilities, particularly in the technical, scientific and engineering fields. A year later about seventy students, attracted by a pointer programme in co-operative education, attended the institution's first classes. Full university powers were conferred by a 1959 Act and the next year the University awarded its first degrees. St. Jerome's College, a century-old Roman Catholic educational institution, federated with the University of Waterloo in 1960 and within the next year Renison (Anglican), St. Paul's (United Church) and Conrad Grebel (Mennonite) became affiliate colleges of the University.

Basic Knowledge – The definition of line-breaking problem

We can now finally give the definition to line-breaking problem.

Legal breaking point: If an item x_i satisfies any rules listed below, we define the number i as a legal breaking point.

1. x_i is a penalty item and the corresponding p_i is not infinity;
2. x_i is a glue item and x_{i - 1} is a box item.

line-breaking problem: For a paragraph x_1, … , x_m, the line-breaking asks for choosing legal breaking points 1 <= b_1 … b_k = m and breaking the paragraph into k lines.
The chosen legal breaking points must contain the index of any penalty item whose penalty point is -infinity.

The K-P Algorithm – Adjustment Ratio

We need a method to evaluate the how bad or how good a line-breaking choice is. The method used in K-P is *Adjustment Ratio*.

Adjustment Ratio: For a line, The actual length is defined as L_i, The stretchability Y_i is defined by summing all of the y_i in the line. The shrinkability Z_i is defined by summing all of the z_i in the line. The ideal length for $l\ \ i$.
The *Adjustment Ratio* for the line r_i is defined as:

$$r_i = \begin{cases} 0 & L_i == l_i \\ (l_i - L_i)/Y_i & l_i > L_i \\ (L_i - l_i)/Z_i & l_i < L_i \end{cases}$$

We suppose that the r_i
should not stretch too muc
orshouldnotshrinktoomuch.

The K-P Algorithm – First-fit Criterion

We want the line-breaking problem to find breaks such that lr_il < 1 in each line, with the minimum number of hyphenations subject to this condition.

This method is

1. Line split locally,
2. Not use every possible values, therefore sometimes no result.
3. Quick

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.

generated by first-fit

The K-P Algorithm – Best-fit Criterion

Each line has been broken without looking ahead to the end of the paragraph
and without going back to reconsider previous choices, but this time each break was chosen so as to minimize that line.

$$\beta_j = \begin{cases} \infty, & \text{if } r_j \text{ is undefined or } r_j < -1; \\ 100|r_j|^3, & \text{otherwise.} \end{cases}$$

This method is

1. Line split locally,
2. Result is a little bit better
3. Quick

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.

generated by best-fit

The K-P Algorithm – Optimum-fit Criterion

It is globally optimum in the sense of having fewest total 'demerits' over all choices of breakpoints, where the demerits assessed for the line j are computed by the formula:

$$\delta_j = \begin{cases} (1+\beta_j+\pi_j)^2+\alpha_j, & \text{if } \pi_j \geqslant 0; \\ (1+\beta_j)^2-\pi_j^2+\alpha_j, & \text{if } -\infty < \pi_j < 0; \\ (1+\beta_j)^2+\alpha_j, & \text{if } \pi_j = -\infty. \end{cases}$$

This method is

1. Line split globally
2. The best but
3. Time consuming.

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.

generated by optimum-fit

The K-P Algorithm – Optimum-fit Algorithm

For a paragraph with n legal breaking points, there will be 2^n possible breaking choices for optimum-fit.

How to overcome that?

The intuition is that there is no need to test all possible breaking choices. For instance, if a line can contain 400 characters, breaking at the 10th character could hardly be the best choice.

Therefore, we could add additional conditions to each line, e.g. we could limit every line with the |r_j| <= 1. and relex conditions when no possible result.

The K-P Algorithm – Optimum-fit Algorithm

With the condition for each line, we also need an algorithm to detect all possible breaking choices.

The K-P algorithm adopts a dynamic programming(DP) method for that. Now we define the function DP(j, k) as the minimum demerit value in total if the line L_k breaks at x_j, and we define the Cost(i, j, k) as the potential cost if we break the L_k at x_j and L_{k-1} breaks at x_i.

$$DP(j,k) = \min_{i=0}^{j-1}(DP(i,k-1) + Cost(i,j,k))$$

The K-P Algorithm – Optimum-fit Algorithm

Assume that we are trying to split:

The quick brown fox jumps over the lazy dog

with each line containing at most 15 units.

1. For the first glues after "The" and "quick", the r_1 will become too large therefore we will not break it.
2. Now the algorithm will check the glue after "brown" and notice that if the line breaks here, the condition of r_1 will be satisfied.
3. it will list the glue item as a possible candidate and go on.
4. Repeat until the end.

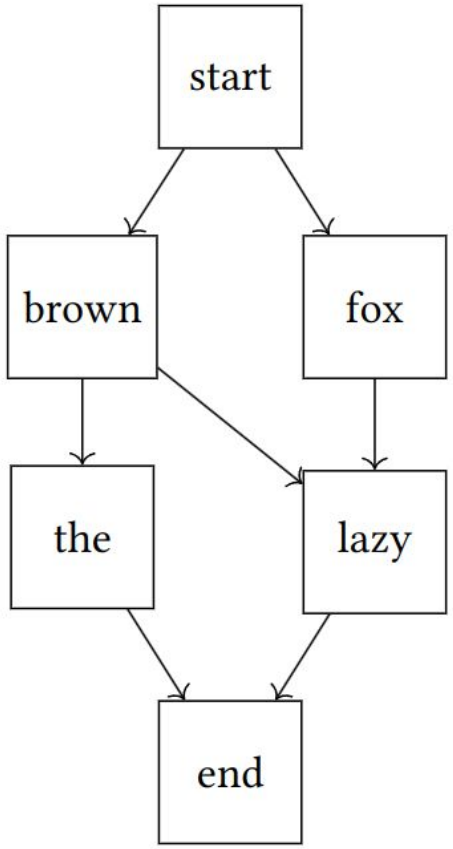


Figure 2: The tree of possible candidates

The K-P Algorithm – Three phases

The algorithm will use three constant values: pretolerance, tolerance, and threshold to indicate the limit of each phase.

In Phase 1, only the glue items and penalty with p_i = -infinity item are considered. The best-fit criterion will be used. If no possible result goto Phase 2.

In Phase 2, every possible legal breaking point are considered, and the optimum-fit criterion will be used. If no possible result goto Phase 3.

In Phase 3, repeat Phase 2 with an emergency stretch allowing stretching or shrinking.

Improvements – re-investigate the K-P algorithm

Although the K-P algorithm is used for the
line-breaking problem, we can rely on our description to conclude that the K-P algorithm is *directly* applied on the box-glue-penalty system.

the K-P algorithm can be applied to other problems so long as they can be translated to the box-glue-penalty system.

Like pagination and micro and macro interaction.

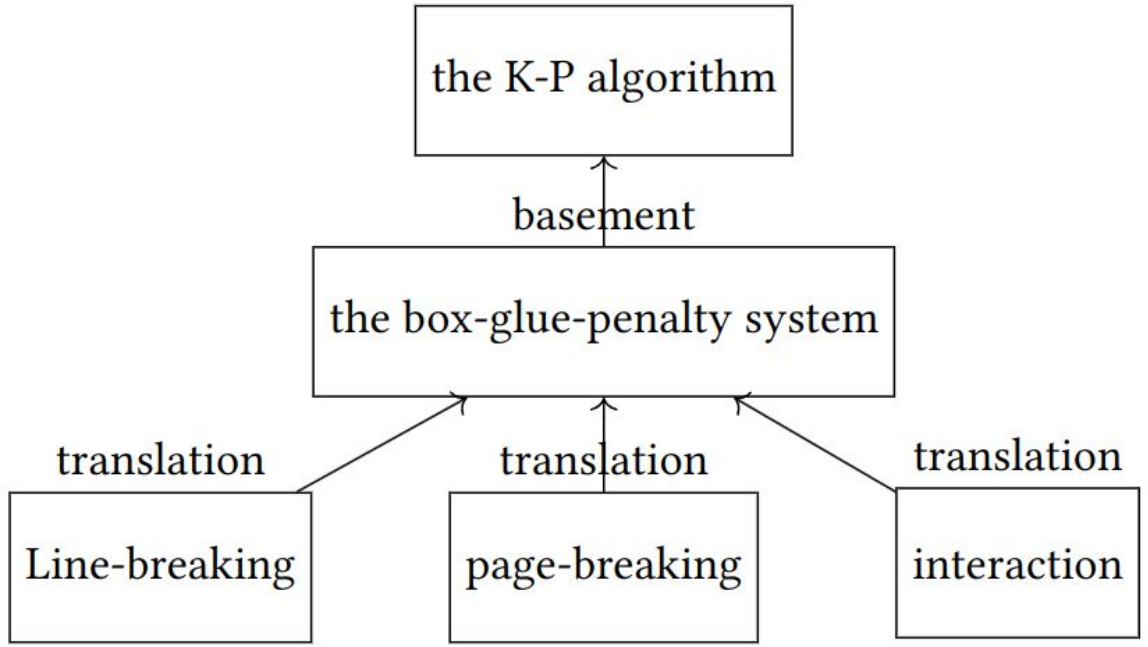


Figure 3: The relationship between the K-P algorithm, the box-glue-penalty system, and problems

Improvements – Line-breaking and Page-breaking

In the Tex system, it uses the K-P algorithm to handle line-breaking, but uses greedy algorithms for pagination.

Back in the 1980s, the computer at that time is slow and could not even store the whole article. therefore perform typesetting page-by-page is the only possible choice.

But now, things changed.

The inherent similarity between link-breaking and page-breaking make it possible to apply the K-P algorithm on the latter problem.

Item	Line-breaking	Page-breaking
Box	character	Line
Glue	space between characters	space between lines
Penalty	allowable breaks between characters	allowable breaks between lines

Improvements – The equivalent problem

Apart from similarity, there still exist some differences between the line-breaking and page-breaking problem.

The first difference is an equivalent requirement.

All columns of a page and in a double spread (facing pages in the output document) need the same space width. This is to hide the width changes from readers and avoid noticeable differences from readers.

This have been solved by following methods.

Improvements – The equivalent problem

For any breaking point b, we will do as follows:

1. If the breaking point is at the end of a page, the next column could generate lines with flexibility. This flexibility is reflected by adjusting the ideal length with the *variation* variable.
   For a given ideal page length C_i, we will generate the fittest page with an ideal length with the *variation* value of (0, -baselineskip, baselineskip), respectively.

2. If the line does not break at the end of a page, the *variation* value would be determined by the previous column.
   Therefore the next page-breaking point is determined and will not be searched for optimal results.

Improvements – Float placement problem

Another difference is that page-breaking needs to place float properly according to the texts.

Breaking texts with floats into different pages is different from pure texts. That is because the float forms an independent input stream for the text.
Floats from one input stream are only expected to appear in the sequential order, but the order from different streams are usually not restricted.

This problem has been partly solved as presented.

Improvements – Float placement problem

We can provide several rules about floats to limit possible float placement choice.

1. Floats will be placed in order of their first call-out.
2. A float is better to not appear before the occurrence of its first call-out occurs.
3. If a float needs to appear before the first call-out, it can only appear on the page before, so long as it is still visible from the first call-out.

And add the float placement to the DP function:

$$DP(i,j,k) = \min_{a=1}^{j} \min_{b=1}^{k} (DP(i-1,a,b) + Cost(i-1,i,a+1,j,b+1,k))$$

Improvements – Line-breaking and Layout

Previous works, including the K-P algorithm, all assume that the best line-breaking result
could lead to the best layout.
Therefore, they adopt the result of the line-breaking algorithm, and perform page-breaking algorithms on that.

But more and more have challenged this assumption, stating that less-optimal results in the K-P algorithm could lead to a better overall layout.

We will present some related solutions.

Improvements – Line-breaking and Layout

The solution integrates the micro- and macro-typography by delaying the definite choice of line breaks and offering a flexible glue item set to the pagination algorithm.

In micro-typography, it use the K-P algorithm to detect several possible line results, optimal and sub-optimal, recording their minimum lines, optimum lines and max lines.

In macro-typography, it enode the paragraph as a glue item glue(opt, opt-min, opt-max) and let the K-P decides each value of the r_i.

For each decided r_i, it will go back to micro-typography to find the suitable result.

Future

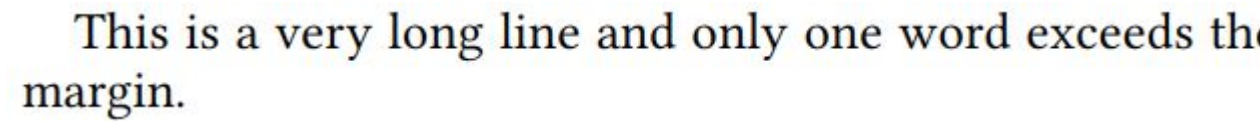
1. **Practical Pagination**
   To the best of our knowledge, the Tex system is still using the greedy algorithm as the default pagination algorithm.

2. **Page-limited Pagination**
   It is a common requirement for writers to limit papers to certain pages, and the page limitation should be considered in the pagination.

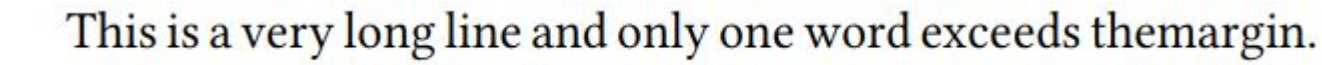
3. **Irregular Float Placement**
   The method related to the K-P algorithm assumes that all floats will be regular, i.e. it will occupy the whole line of text.
   But the float may be surrounded by texts, and the float shape itself is irregular.

Bonus – Use it at your own risk!

```
{\large
This is a very long line and only one word exceeds the margin.
}
```


This is a very long line and only one word exceeds the
margin.

```
{\large
This is a very long line and only one word exceeds the\penalty10000 margin.
}
```


This is a very long line and only one word exceeds themargin.

Thank you