Eurographics Conference on Visualization (EuroVis), Posters Track (2016)
T. Isenberg and F. Sadlo (Guest Editors)

# InVITe - Towards Intuitive Visualization of Iterations over Text

H. Lamqaddam[†1,2] and J. Aerts[1,2]

[1]Visual Data Analysis Lab (VDA-lab), ESAT/STADIUS, KU Leuven, Belgium
[2]iMinds Medical IT, KU Leuven, Belgium

**Abstract**

*With InVITe, we are working towards intuitive visualization to support review of iterative modifications on text documents. In order to accomplish this, we perform simple matching of text snippets between the two versions of text, across a large range of parameter settings. Next, an overview graphic indicating the effect of parameter space on the output allows the user to select those combinations that are of interest. Finally, such selection will display an alluvial diagram with annotations and covering different resolutions.*
*With this tool, co-authors can keep an overview of changes made, both structural and local.*

Categories and Subject Descriptors (according to ACM CCS): Information Interfaces and Presentation (e.g., HCI) [H5.2]: User Interfaces—Graphical user interfaces (GUI); Document and Text Processing [I.7.0]: General—

## 1. Introduction

Preparation of text documents is at the heart of a vast number of professional disciplines, and includes for example legal documents, analysis reports, meeting minutes, project proposals; the list is endless. In many cases, the authors will iterate over that text to ensure that the message contained in it is conveyed in an optimal way. In a collaborative setting, unfortunately, it can be difficult to keep track of what has actually changed.

Consider a typical collaborative process between a postgraduate student and his/her promoter while preparing a scientific paper. The student will prepare a first draft, which is reviewed by the promoter. The latter has several remarks, including small changes as well as substantial restructuring of the text itself. The postgraduate student implements these suggestions using the "track changes" feature in Microsoft Word. Although this allows the promoter to identify the parts that have been changed, this approach is crude and has many shortcomings. For example, when a section of text (e.g. paragraph or chapter) is moved to another position, it is indicated as both a deletion and insertion, rather than a translocation. In addition, smaller changes are embedded and hidden in these large ones. It is also not possible at the moment to have an overview of the evolution of the text across multiple versions: which parts have been stable, which were moved, which had internal changes, etc.

A wide range of solutions have been described before for the visualization of text. A comprehensive overview can be found at http://textvis.lnu.se/ [KK15]. Most of these are however concerned with the representation of a single text (e.g. [WV08], [VWF09], [RGP*12]), overviews of text/document corpora (e.g. [SSNR14], [FGM05], [CL09]), or issues like topic extraction and/or evolution (e.g. [GJG*15], [LKC*12]). Conceptually, text documents are of the same type as software code and genome sequences, as each of these also consists of a sequence of characters. Visualizations have been developed for investigating programming code (e.g. [VTVW05], [BK01], [ESSJ92]) as well as genomic structural variations (e.g. Circos plot as used in [ZBJL*10] and [MGB*14], [PT03]), but neither is fit for the task described here. Fry's visualization of the evolution of The Origin of Species by Charles Darwin (http://fathom.info/traces/) does indicate how a text changes over subsequent iterations, but does not allow for looking at this at different resolutions.

Overall, it is clear that the current approaches do not take into consideration the specific tasks of the writer/reviewer: in the (re)writing process, one typically switches between considering the overall structure of the text, and investigating smaller changes such as spelling (e.g. British vs American English).

With InVITe, we developed a tool able to look at these different resolutions of text differences, giving the (co)authors the ability to investigate small and large changes in context.

## 2. Approach

### 2.1. Algorithm

The interactive InVITe visualization relies on a simple text analysis approach, inspired by the bl2seq algorithm for DNA sequence

† Corresponding author: houda.lamqaddam@kuleuven.be

comparison [CCA*09]. The original text is considered the reference and divided into atoms with a pre-defined window size *w*. The algorithm then scans the second text in search of these atoms. This match does not have to be perfect as we allow an error rate *e* between the reference atom and its match. Using large values for window size *w* and error *e* favours visualization of large structural changes in the text (e.g. new, deleted or translocated sections), whereas small values favour small local changes such as spelling or word choice. Our tests indicate that running 1,600 combinations of the *w* and *e* parameters on two versions of a 216-page document takes 21 seconds on a 2.2 GHz Intel Core i7 Mac laptop.

### 2.2. Visualization

As these parameters *w* and *e* have a significant impact on the resulting plot, InVITe provides an overview of this parameter space as indicated in Figure 1. This allows the user to choose the granularity of the returned visual, corresponding to the task that the user wants to perform. Selecting a combination of *w* and *e* results in the alluvial diagram as presented in Figure 2. This plot consists of several parts. Part A represents the reference text, indicating each chapter in a different colour and including the section title if the text is written in Markdown syntax. A diagonal line across the box indicates the position within that chapter; an approach regularly used in comparing genomes between species, and described in Figure 12 of [HMB*04]. Part B shows the new text version, in its rearranged state. Chapters of origin and rearrangements within them can be easily identified using the colour encoding and diagonal line. In addition, the marks on the right indicate stability, showing where breaks in the alluvial diagram would appear if the user were to choose a smaller window size *w* and/or lower allowed error rate *e*. Hovering the mouse over a section in either the original or new version will show the underlying text. An alluvial diagram (C) connects the two text versions. Clicking on a band (C) will open a side-by-side view, zoomed into that section, using more stringent parameter settings and therefore splitting the text further into subsections.

## 3. Conclusion

With InVITe, we developed a tool for investigating different iterations of a text, at different resolutions. InVITe is a work in progress and future work will include refinement of the mapping algorithm (e.g. by adapting the bl2seq program itself [CCA*09]), live editing of these texts with immediate modification of the plots, and the comparison of more than 2 versions.

Overall, we believe that the InVITe tool can greatly support authors in reviewing different iterations of medium- to large-sized text documents.

## Acknowledgements

The work presented here is supported by IWT/SBO grant ACCUMULATE 150056 and iMinds Medical IT.

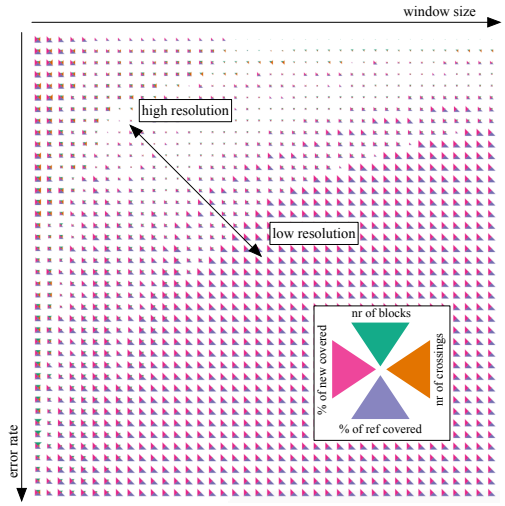


**Figure 1:** *Matrix covering parameter space for window size* w *and error rate* e*, and indicating their effect on the resulting blocks. Each cell corresponds to a different instance of the graphic in Figure 2. Plots in the lower right corner indicate that with high* e *and* w *one ends up with a large matches of single blocks; using low* e *and* w *(top left) identifies crossings and multiple blocks.*

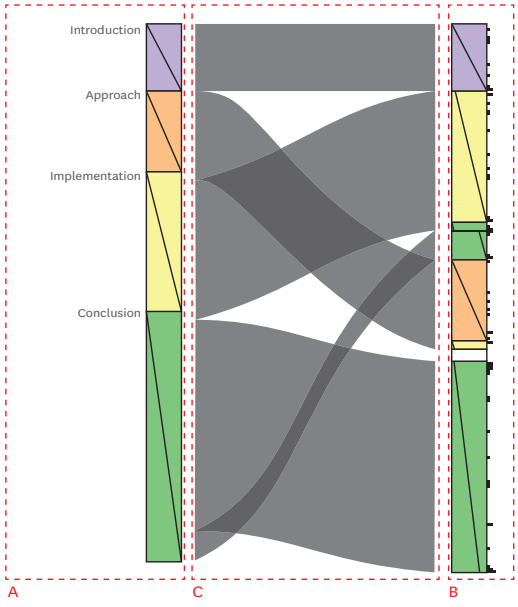


**Figure 2:** *Alluvial diagram showing how two versions of text differ from each other, as well as the sensitivity of each section with changes in parameter settings. For full description, see main text.*

## References

[BK01] BASSI S., KELLER R. K.: Software visualization tools: Survey and analysis. In *Program Comprehension, 2001. IWPC 2001. Proceedings. 9th International Workshop on* (2001), IEEE, pp. 7–17. 1

[CCA*09] CAMACHO C., COULOURIS G., AVAGYAN V., MA N., PA-

PADOPOULOS J., BEALER K., MADDEN T. L.: Blast+: architecture and applications. *BMC bioinformatics 10*, 1 (2009), 1. 2

[CL09] CULY C., LYDING V.: Corpus clouds-facilitating text analysis by means of visualizations. In *Human Language Technology. Challenges for Computer Science and Linguistics*. Springer, 2009, pp. 351–360. 1

[ESSJ92] EICK S. G., STEFFEN J. L., SUMNER JR E. E.: Seesoft-a tool for visualizing line oriented software statistics. *Software Engineering, IEEE Transactions on 18*, 11 (1992), 957–968. 1

[FGM05] FORTUNA B., GROBELNIK M., MLADENIC D.: Visualization of text document corpus. *Informatica 29*, 4 (2005). 1

[GJG*15] GAD S., JAVED W., GHANI S., ELMQVIST N., EWING T., HAMPTON K. N., RAMAKRISHNAN N.: Themedelta: dynamic segmentations over temporal topic models. *Visualization and Computer Graphics, IEEE Transactions on 21*, 5 (2015), 672–685. 1

[HMB*04] HILLIER L. W., MILLER W., BIRNEY E., WARREN W., HARDISON R. C., PONTING C. P., BORK P., BURT D. W., GROENEN M. A., DELANY M. E., ET AL.: Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature 432*, 7018 (2004), 695–716. 2

[KK15] KUCHER K., KERREN A.: Text visualization techniques: Taxonomy, visual survey, and community insights. In *Visualization Symposium (PacificVis), 2015 IEEE Pacific* (2015), IEEE, pp. 117–121. 1

[LKC*12] LEE H., KIHM J., CHOO J., STASKO J., PARK H.: ivisclustering: An interactive visual document clustering via topic modeling. In *Computer Graphics Forum* (2012), vol. 31, Wiley Online Library, pp. 1155–1164. 1

[MGB*14] MONCUNILL V., GONZALEZ S., BEÀ S., ANDRIEUX L. O., SALAVERRIA I., ROYO C., MARTINEZ L., PUIGGRÒS M., SEGURA-WANG M., STÜTZ A. M., ET AL.: Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nature biotechnology 32*, 11 (2014), 1106–1112. 1

[PT03] PEVZNER P., TESLER G.: Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome research 13*, 1 (2003), 37–45. 1

[RGP*12] RIEHMANN P., GRUENDL H., POTTHAST M., TRENKMANN M., STEIN B., FROEHLICH B.: Wordgraph: Keyword-in-context visualization for netspeak's wildcard search. *Visualization and Computer Graphics, IEEE Transactions on 18*, 9 (2012), 1411–1423. 1

[SSNR14] SIIRTOLA H., SÄILY T., NEVALAINEN T., RÄIHÄ K.-J.: Text variation explorer: Towards interactive visualization tools for corpus linguistics. *International Journal of Corpus Linguistics 19*, 3 (2014), 417–429. 1

[VTVW05] VOINEA L., TELEA A., VAN WIJK J. J.: Cvsscan: visualization of code evolution. In *Proceedings of the 2005 ACM symposium on Software visualization* (2005), ACM, pp. 47–56. 1

[VWF09] VIEGAS F. B., WATTENBERG M., FEINBERG J.: Participatory visualization with wordle. *Visualization and Computer Graphics, IEEE Transactions on 15*, 6 (2009), 1137–1144. 1

[WV08] WATTENBERG M., VIÉGAS F. B.: The word tree, an interactive visual concordance. *Visualization and Computer Graphics, IEEE Transactions on 14*, 6 (2008), 1221–1228. 1

[ZBJL*10] ZEITOUNI B., BOEVA V., JANOUEIX-LEROSEY I., LOEILLET S., LEGOIX-NÉ P., NICOLAS A., DELATTRE O., BARILLOT E.: Svdetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics 26*, 15 (2010), 1895–1896. 1