Les Cahiers du GERAD ISSN: 0711–2440

The progressive visualization, a new tool for analyzing the writing process

H.-S. Bécotte, G. Caporossi,
A. Hertz

G–2015–141

December 2015

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

The progressive visualization, a new tool for analyzing the writing process

Hélène-Sarah Bécotte [a]
Gilles Caporossi [b]
Alain Hertz [a]

[a] *GERAD & Polytechnique Montréal, Montréal (Québec) Canada, H3C 3A7*

[b] *GERAD & HEC Montréal, Montréal (Québec) Canada, H3T 2A7*

helene.becotte@gerad.ca
gilles.caporossi@gerad.ca
alain.hertz@gerad.ca

December 2015

Les Cahiers du GERAD
G–2015–141

Abstract: Writing is a complex process and it is difficult to know how to write well in order to make a good text. Many fields and techniques are combined to analyze how good writers operate compared to poor writers. Today it is possible to record a writing session on a computer and to have access to all the different operations that are made and when it happened. The problem with these files is that they contain a lot of data that is not well suited to be analyzed by humans without preprocessing. The goal of this research is to automatize that preprocessing step and help the searchers in their analyses of the writing process using visualization techniques.

Résumé : Écrire un texte de qualité est un processus complexe. Plusieurs domaines d'études se penchent sur les possibilités d'analyse des pratiques d'écriture des bons scripteurs par rapport aux moins bons afin de comprendre les stratégies d'écriture gagnantes. Grâce à la technologie actuelle, il est possible d'enregistrer l'ensemble des opérations effectuées lors de la rédaction d'un texte sur ordinateur, ce qui en facilite l'étude. Par contre, les fichiers contenant l'information contiennent beaucoup de données et sans pré-traitement, elles ne sont pas appropriées pour l'analyse humaine. Nous proposons une nouvelle méthode de visualisation des données pour permettre de faciliter et d'automatiser une partie de l'analyse du processus d'écriture.

1 The challenges of understanding the writing process

Writing a good text is not easy and requires a specific talent [24]. It has also been suggested that the skill of being a good writer improves with time and practice [29].

Nowadays, many softwares have been created to improve writing as an administrative task but none is oriented on literary creation [24]. Their use helps writing tasks in a business environment but they don't transform their users into skilled or good writers. To understand the great authors' writing creation mechanisms and strategies, psychologists, linguists and didacticians studied notorious authors such as Flaubert, Proust, Valéry and Zola [14]. The science of studying the writing process is called genetic criticism and its goal is to understand what kind are the different revisions made along the process that makes the text good at the end [3].

Many pattern assumptions have been studied already in the writing process. Starting from 1981, researchers suggested that the level of the writer's proficiency influenced the nature and the frequency of revisions. It seems that less skilled writers don't revise much and their modifications are mainly about syntax and punctuation [8] while good writers revise more and change the meaning of the text often over larger part of their written production [17].

We don't know much about the text production practice in history and the writing process is far from being known in detail [24]. With the progress made in software technologies, it is possible to record each operation made by a writer and therefore to access detailed information about this persons's writing process [27]. The files containing that information are called log. This data stored in those files is big and complex and is not appropriate for a human to analyze it without preprocessing [10].

Only a few tools exists to help analyzing the writing process. Two different aspects makes this process difficult to tackle [33]. First, the process can be modelled with many dimensions and secondly, each writing process is fundamentally different. Some authors already suggested that automatic analysis would be useful in linguistics. In 1997, machine learning and statistical pattern recognition, two data mining techniques, were used to automatically extract implicit or unknown facts from it [13].

2 The role of visualizations for pattern recognition

When tackling big amount of data, using visualizations is a great way to explore and analyze a given situation [38]. Visualizations can help identifying patterns, structures among data, irregularities and relationships over a specific timespan [30]. Their goal is to let the human eye find underlying structures [5]. After the data analyst understood and gained insight into the visualized data, he or she can draw conclusions from the discovered patterns [16]. This analysis method has its flaws: the human ability's to process a lot of data is limited [28] and sometimes researchers tend to see relations between random and unrelated items, which is called apophenia [7]. In order to compensate those interpretation bias, "being able to derive scientific insight from data increasingly depends on having mathematical and perceptual models to provide the necessary foundation for effective data analysis and comprehension" [31].

Displaying data with graphic techniques may not be sufficient to help the user finding patterns [35]. The way people interact with the visualization tool and perceive the displayed data influences how the users understand the relationships and potential patterns [35]. Therefore, before creating a new visualization, it is important to think about the human factor and about ways to improve its effectiveness.

A lot of research has been done to understand the cognitive support that visualizations give to assist the human brain while trying to visually understand data [35]. In order to create good visualizations, it is essential to consider two related concepts which are data representation and data presentation. When choosing the graph or chart type that illustrate the data's variables, it's considered to be the data representation step [21]. The delivery format, the overall appearance and design, including the choice of colour and the interactive features are part of the data presentation [2].

2.1 Data representation

The data representation is all about choosing the right physical way to represent data [21]. In order to create a visual version of the data, it is essential to think about what kind of variables will be represented. It is also important to think about the goal of the representation: whether it will be used for explorative analysis, confirmative analysis or simply the presentation of the results [2].

If time is a variable, the fact it is considered as an absolute number or if it is interpreted relatively changes the way it can be represented. When it's absolute, position and lenght are used. In the case of a visualization in two dimensions such as what is displayed on a paper sheet, the time is usually mapped with a map or graph with two axis, time being displayed on one of them [2].

In the case where time can be interpreted, visual variables are used such as color and transparency to illustrate the duration between two events for example [2].

The other type of variables can either be represented on one of the axis or with shape or color to differenciate them [21].

Data visualisations are traditionaly static [21]. In order to properly illustrate how the data interacts, it is important to show change over time [4]. Because static visualizations are in two dimensions, it is complicated to make them represent efficiently problems that have more than two dimensions [30]. Consequently, one of the best ways to visualize spatiotemporal changes is to animate the data [38]. Spatiotemporal representations should be based on interactive sources that could allow the user to manipulate the data to allow a deeper understanding [32]. These functionalities help the user to be involved into a dynamic problem-solving methodology and go further in their analysis. Allowing the manipulation of variables and parameters is a way of showing many dimensions within a single display and to facilitate the combination of different variables when exploring visualized data [21].

To help the user investigate data, some authors suggested to include tasks that may be performed on the visualized data (Shneiderman 1996, as cited by [2]). Adjusting the view with the possibilities of having an overview and/or to zoom to change the degree of accuracy allows the user to see either the entire dataset or a more specific part of it and can be useful to use the data to achieve different goals. To filter and extract variables allows the user to remove unwanted information or to extract data with query parameters like sorting the data, selecting a specific temporal window, a selected spatial or geographic space or other statistical queries. Sometimes it is essential to gain specific insights to confirm or infirm a correlation, pattern or relationship within the data.

When the data represents multidimensional information with a time-varying phenomena, it is essential to access the different structures and underlying layers. Even if the use of the different tasks is possible, it is necessary to have a mathematical proof of those patterns and relationships. Combining the visualization with mathematical data analytics is necessary for understanding the structures and relationships found [31].

2.2 Data presentation

That data representation is not the only aspect involved when creating visual support to describe and understand data. The overall ergonomics of the visualization is important to maximize the analysis potential.

The presentation of data involves also the overall design of the visualization tool. While it is more important when the goal of the visualization is to present some results, some features are worth considering in an explorative-orientated model [36].

The most important aspect would be the choice of colors. According to Bartram [4], an appropriate combination of shape, symbols, colour, size and position should be chosen because that kind of visual information doesn't require cognitive effort and is rather processed by the preattentive visual system. Displaying a complete combination of visual codes and dimensions increase the analyst's efficiency and speed while looking at the representation. While colors can help illustrate variables, as seen in the data representation section, an other consideration when choosing those colors would be to make the visualization tool accessible to the greater number of researcher by choosing colors while paying attention to how color blind persons (see [2]).

3 Analyzing the writing process

To understand the way we write and the way we create texts is not an easy task [33]. Many factors such as the writer's ability to retain content by memorizing it, the level of proficiency in writing in that specific language by the writer and the specification of the task for example influence the writing process. Many researchers specialized in genetic criticism to try to understand the variables that influence the writing process [26]. Another way of analyzing this is by looking, step by step, at the operations made by the writer.

Basically, only two kind of operations are made by a writer. A character and either be added or deleted. In the literature, adding a keystroke is called an *insertion* and deleting it, a *deletion* [9].

The writing process is characterized by its multi-dimensionality [10]. There are two basic dimensions: chronology and spatiality. The chronology represents the order of the operations in the text as they have been made in time. Therefore, two operations can't have the same chronology position. The spatiality concerns where the operations have been made geographically in the text.

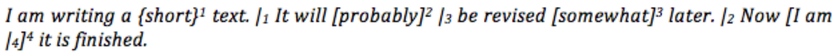
It is essential to know that this is not a mathematical description of the writing process data. Visualizations are a tool that allow humans to understand data. Dimensions are therefore different. The fact that the state of the text changes with time has to be represented in order to help humans to understand the data. This third dimension, that would be called temporality is more of a visualization dimension than a mathematical dimension. The chronology dimension is in a way a simplification of the temporality. The operations are made chronologically in time but the overall state of the text changes as the writer inserts or delete keystrokes and this is those different states of the text that are precisely studied in genetic criticism [23].

The techniques used actually are almost all related to visualization. The writing process has been represented in many different ways by linguists in order to understand the different revisions made in a text and the general writing process.

There are four main types of visualizations used.

The first one, which is dynamic, consists of the visual reproduction of the operations made in the text. It is somehow a movie representing all the operations made, in their temporal order. However, even if this representation is animated, it is impossible for the user to perform any task on the data [20].

The linear representation is used with different softwares and is somehow a traduction of the log file which is the recording file of the writing process. It can be read similarly to a text. The dimension used in this visualization is the spatiality with some insights about the chronology of the operations. The following example was created by Kollberg [22]:

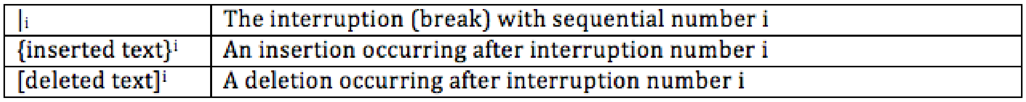


I am writing a {short}^{1} text. $|_1$ It will [probably]^2 $|_3$ be revised [somewhat]^3 later. $|_2$ Now [I am $|_4$]^4 it is finished.

The final text would be: *I am writing a short text. It will be revised later. Now it is finished.*

The notation of this specific linear representation is explained in the table below.

Table 1: S-notation's symbols



$|_i$	The interruption (break) with sequential number i
{inserted text}^{i}	An insertion occurring after interruption number i
[deleted text]^i	A deletion occurring after interruption number i

The GIS representations are actually Geographical Information Systems. Geographical data shares some similarities with the writing process data. The operations are displayed on a graphic with two axis. One represents the temporality dimension and the other one is the spatiality. In this case, the spatiality is the position of the operation as it is recorded in the log file. Many operations can share the same spatiality number [26]. The following example of the GIS representation was made by Leijten and Van Waes [25]:

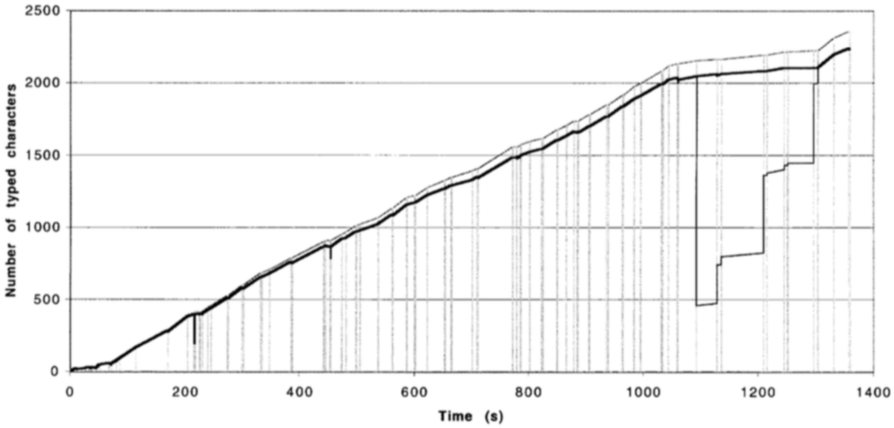


Figure 1: The GIS representation

Finally, the graph representation allows the user to see the dynamic aspect of the writing process [9]. This visualization is based on graph theory. A graph is a visual representation that is made of points tied together with lines [6, p.1]. The points are called nodes and the lines are called edges. Two nodes are said to be adjacent or neighbours if they are linked with an edge [19]. The degree of a node is the number of its neighbours [34]. A graph $G = (V, E)$ is made of a non empty set of V nodes and of a set of E edges. The chronology is the main dimension used in this visualization [9]. It is also possible to see the spatiality. However it is not represented the same way than in the GIS representation. The edges between the grouped operations follows this nomenclature: the chronological order is represented by a solid line and the spatial order with a dotted line; the order in the final text is represented by the color red and is black otherwise. The link between an addition node and its deletion is blue. Figure 2 is an example of graph representation [10]:

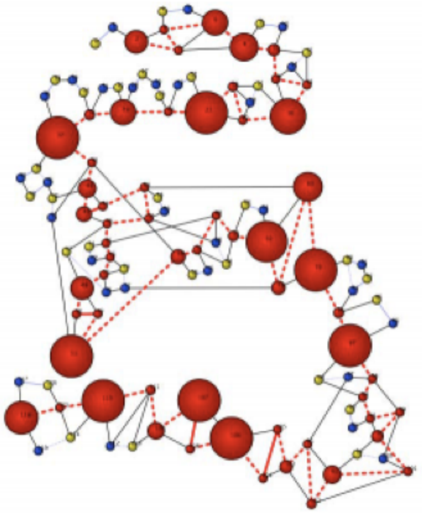


Figure 2: The graph representation

In their software, it is possible, by pointing the mouse to a node, to know its content. It is therefore possible to gain precise insight about the writing process [9].

4 The progressive representation

The progressive representation aims to be an integrative solution that allows the writing process to be displayed in 3 dimensions. The proposed solution is to merge the advantages of the GIS visualizations with the best aspects of the graph representation. Moreover, as the goal is to represent the writing process in a way that captures the temporal aspect, the representation will be dynamic so it is possible to interact with it.

This new model is inspired by both the GIS and the graph representation. The GIS have the advantage of representing well the interaction between the chronology and spatiality of the writing process [26]. Their weakness is to show only a general view of the process, it is impossible to know what has been written, what is the context and what are the exact operations that have been made [9]. As the position used is the one recorded by the log file at the moment the revision was made, it is impossible to know which part of the text is therefore modified by an operation.

The graph representation is helpful to understand the link between many operations and their chronology [10]. However, it is difficult to figure out exactly what is in each node and the temporality of the operations.

The visualizations inspired by the GIS are good at showing the process as a whole and emphazise on temporality and spatiality. On the other hand, the graph representation also allows the user to see the process as a whole but it is possible to gain insight on detailed operations as well. The dimensions used are chronology and spatiality.

The graph is included in a quadrant with the spatiality on the abscissas axis and the temporality on the ordinate. Unlike the GIS, this representation is read from top to bottom to mimic the way we read texts. Another difference is the position shown. In the GIS, the position of the operations is relative and corresponds exactly to the position in the log file. If a writer rewrites the same text portion 5 times, there will be 5 insertions done at the same place on the abscissa axis and 4 deletions.

In this representation, a new position is introduced. Its a hybrid between the relative and the absolute that have been described previously. Every new insertion will have a proper position on the abscissa axis. If an insertion is deleted later on, that deletion will have the same position as that insertion. Only two operations (insertion and deletion) can share the same abscissa value. The highest number of this position is equivalent to the number of keystrokes inserted (letter, space, number) plus the number of pauses of a length greater than a threshold p. Even if a part of the text is deleted, it stays on the visualization. Any part of the text that replaces the deleted part is necessary written physically right after as it is represented on the linear representation.

Contrary to the graph representation where each node was composed of a set of continuous insertions, each sole operation is considered to be a graph node. If two operations have been neighbours, at one point or another during the process they then would be linked with an edge. The edges nomenclature follows the one used in the graph representation. The chronological order is represented by a solid line and the spatial order with a dotted line. The order in the final text is represented by the color red and is black otherwise. The link between an addition node and its deletion is blue [9].

The dynamic aspect of a visual representation is very important [1]. In this model it comes from the zoom in and zoom out property. This function allows the user to explore the details of the process as well as understanding it's dynamic as a whole. Instead of having to use two separate representations to get access to that information, the researcher only needs one, which accelerate its analysis process. As mentioned previously, interacting with data allows the human brain to detect better the patterns and understand the relationship between the variables [21].

It is an extensive version of the graph representation by Caporossi and Leblay [9] where the nodes of the graph are made of a characters string. Sequences of keystroke are merged together to form an entity involved in the conception of the text, which is represented by a node.

Each node represents a keystroke operation. It is linked to the previous temporal operation and to the next one with an edge. This node is also linked to the previous chronological operation and to the next one with an edge. It is possible that the spatial and the chronological neighbours are the same. Their position in the quadrant is exactly (x, y) as x being the spatial position, corresponding to its position number while y is the chronological position, corresponding to the time the operation have been made in the log file.

The following example shows the progressive visualization when used on a precise level.

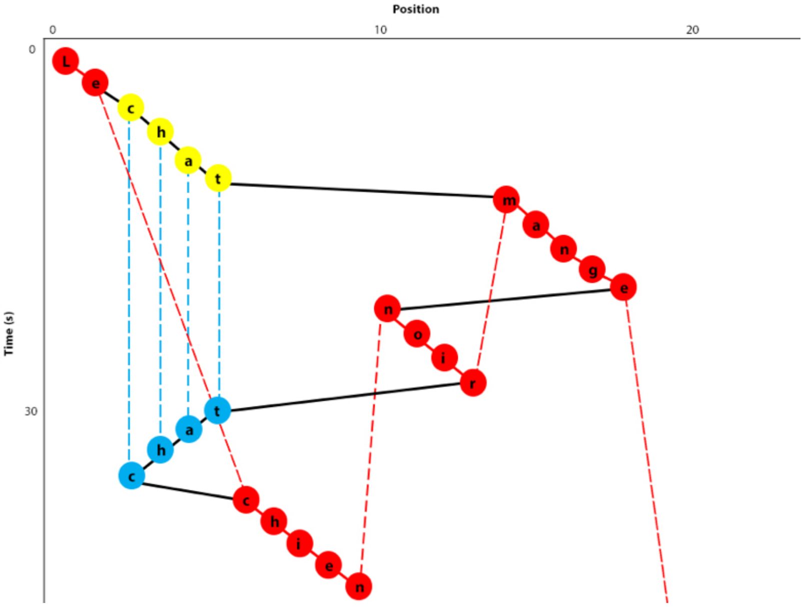


Figure 3: The progressive visualization

The final version of the text can easily be read by following the red nodes and edges. The text would be *Le chien noir mange.*

The writer chronologically wrote *Le chat mange* which can easily be followed by reading the operations along the black straight line. The writer then came back in space and inserted the word *noir* so the text at that moment would be *Le chat noir mange.* The writer then deleted the word *chat* by erasing the keystrokes in a reverse order, most likely using the *backspace* key. After these deletions, the text was *Le noir mange.* Finally, the writer added the word *chien* and the final text corresponding to this part of the writing process is *Le chien noir mange.*

5 Conclusion

The creation of this new representation had two purposes. The first one is to allow the researchers to have a better tool to help them analyze visually the writing process by representing more dimensions. The second one is to structure the writing process data so it's easier to analyze with data analysis techniques. Combining mathematical models with the human inputs gained with the use of visualizations is the key to maximize the quality of the results found [31].

References

[1] F.A. Abukhodair, B.E. Riecke, H.I. Erhan, and C.D. Shaw. Does interactive animation control improve exploratory data analysis of animated trend visualizations? Proc. SPIE 8654, Visualization and Data Analysis 2013, Burlingame, CA, USA, 2013.

[2] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. Visualization of time-oriented data. Human-Computer Interaction Series, Springer-Verlag London, 2011.

[3] D. Alamargot and J.-L. Lebrave. The study of professional writing: A joint contribution from cognitive psychology and genetic criticism. European Psychologist, 15:12-22, 2009, http://dx.doi.org/10.1027/1016-9040/a000001.

[4] L. Bartram. Perceptual and interpretative properties of motion for information visualization. Proc. 1997 workshop on New paradigms in information visualization and manipulations, pp. 3–7, 1997.

[5] F. Blanchard. Visualisation et classification de données multidimensionnelles. Application aux images multicomposantes. PhD thesis, Université de Reims Champagne-Ardenne, 2005.

[6] A.J. Bondy and U. Murty. Graph Theory With Applications. Elsevier Science, 1982.

[7] D. Boyd and K. Crawford. Critical questions for big data. Information, Communication and Society, 15(5):662–679, 2012.

[8] I. Breetvelt, H. Van Den Bergh, and G. Rijlaarsdam. Relations between writing processes and text quality: When and how. Cognition and Instruction, 12(2):103–123, 1994.

[9] G. Caporossi and C. Leblay. Online writing data representation: A graph theory approach. In Advances in Intelligent Data Analysis X, Lecture Notes in Computer Science Series, 7014:80–89, 2011.

[10] G. Caporossi and C. Leblay. Outils de visualisation de données enregistrées. In Temps de l'écriture: enregistrements et représentation, C. Leblay and G. Capossossi (Eds.), Coll. Sciences du langage : carrefours et points de vue, Academia, pp. 147–166, 2014.

[11] D. Chakrabarti and C. Faloutsos. Graph Mining: Laws, Tools, and Case Studies. Morgan and Claypool Publishers, 2012.

[12] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. Introduction to Algorithms. Massachusetts Institute of Technology, third edition, 2009.

[13] W. Daelemans, P. Berk, and S. Gillis. Data mining as a method for linguistic analysis: Dutch diminutives. Dutch diminutives, Folia Linguistica, XXXI/I-2:57–75, 1997.

[14] C. Doquet. Pour une approche linguistique de l'écriture enregistrée. In Temps de l'écriture: enregistrements et représentation, C. Leblay and G. Capossossi (Eds.), Coll. Sciences du langage : carrefours et points de vue, Academia, pp. 21–42, 2014.

[15] A. Dubey and S. Sharma. Comparison of graph clustering algorithms. International Journal of Computer Trends and Technology, 4(9):3230–3235, 2013.

[16] G. Dzemyda, O. Kurasova, and J. Kilinzkas. Multidimensional data visualization: Methods and applications. Springer Optimization and its Applications Series, Vol. 75, Springer, 2013.

[17] L. Flower and J.R. Hayes. A cognitive process theory of writing. College Composition and Communication, 32(4):365–387, 1981.

[18] S. Fortunato. Community detection in graphs. Physics Reports, 486(3–5):75–174, 2010.

[19] E. J. Henley and R Williams. Graph Theory in Modern Engineering. Academic Press, 1973.

[20] ITEM. Enjeux de recherche.

[21] A. Kirk. Data Visualization: A Successful Design Process. Packt Pub, 2012.

[22] P. Kollberg. S-notation as a tool for analyzing the episodic nature of revisions. In European Writing Conference, 1996.

[23] C. Leblay. Le temps de l'écriture : genèse, durée, représentations. PhD Thesis, University of Jyväskylä, Finland, 2011, http://urn.fi/URN:ISBN:978-951-39-4519-0.

[24] J.-L. Lebrave. Comment écriront-ils? Diogène, 196(4):163–171, 2001.

[25] M. Leijten and L. Van Waes. Inputlog: New perspectives on the logging of on-line writing processes. In Computer Keystroke Logging and Writing: Methods and Applications, K.P.H. Sullivan and E. Lindgren (Eds.), Elsevier, 2006.

[26] E. Lindgren and K.P. Sullivan. The LS graph: A methodology for visualizing writing revision. Language Learning, 52(3):565–595, 2002.

[27] E. Lindgren and K.P. Sullivan. La révision en production écrite enregistrée. In Temps de l'écriture: enregistrements et représentation, C. Leblay and G. Capossossi (Eds.), Coll. Sciences du langage : carrefours et points de vue, Academia, pp. 71–92, 2014.

[28] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A.H. Byers. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011.

[29] E. Midgette, P. Haria, and C. MacArthur. The effects of content and audience awareness goals for revision on the persuasive essays of fifth- and eight- grade students. Reading and Writing, 21:131–151, 2008.

[30] M. Minelli, M. Chambers, and A. Dhiraj. Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses. Wiley Publishing, 2013.

[31] T. Möller, B. Hamann, and R. Russell. Mathematical Foundations of Scientific Visualization, Computer Graphics, and Massive Data Exploration. Mathematics and Visualization Series, Springer, 2009.

[32] F.J. Ohlhrost. Big Data Analytics: Turning Big Data Into Big Money. Wiley Publishing, 2013.

[33] S. Plane, D. Alamargot, and J.-L. Lebrave. Temporalité de l'écriture et rôle du texte produit dans l'activité rédactionnelle. Langages, 177(1):75–88, 2010.

[34] S. Saha Ray. Graph Theory With Algorithms and Its Applications in Applied Science and Technology. Springer India, 2013.

[35] M. Tory and T. Moller. Human factors in visualization research. IEEE Transactions on Visualization and Computer Graphics, 10(1):1–13, 2004.

[36] A. Unwin, C.-H. Chen, and W.L. Härdle. Introduction. In Handbook of Data Visualization, C.-H. Chen, W.K. Härdle and A. Unwin (Eds.), pp. 3–14, Springer, 2008.

[37] A. Vathy-Fogarassy and J. Abonyi. Graph-Based Clustering and Data Visualization Algorithms. Springer Briefs in Computer Science Series, Springer-Verlag London, 2013.

[38] N. Yau. Visualize This: The Flowing Data Guide to Design, Visualization and Statistics. John Wiley & Sons, Inc., 2011.