A Method for Thematic and Structural Visualization of Academic Content

Alexander Amigud[1*], Joan Arnedo-Moreno[1], Thanasis Daradoumis[1,2], Ana-Elena Guerrero-Roldan[1].
[1]Department of Computer Science, Multimedia and Telecommunications Universitat Oberta de Catalunya (UOC), Barcelona, Spain. [2] Department of Cultural Technology and Communication University of Aegean, Mytilini, Greece.
aamigud@uoc.edu // jarnedo@uoc.edu // adaradoumis@uoc.edu // aguerreror@uoc.edu

Abstract— **Academic work: grading student assignments or conducting literature surveys entails extensive reading, which is both a time consuming and cognitively demanding task. The challenge increases proportionally with the increase of volume of the textual content. In this paper, we propose a novel approach to visualization of textual data that depicts information on a continuum (temporal or spatial) allowing inferences to be made about thematic organization of a document and its structure. Our visualization method—termed ThemeTrack—creates a visual map: delineating key themes and tracking their presence throughout the text, highlighting their variations and relationships. It aims to make the review of textual data more efficient. To assess the viability of the proposed approach, a series of experiments were conducted using graduate-level theses and published articles in the peer-reviewed journals. The applications of the proposed method are discussed and the real-word examples are provided.**

Keywords-data analytics; knowledge retrieval; technology-enhanced assessment; text mining; visualization methods for learning

I. INTRODUCTION

Reading and writing are two major parts of the academic work. Content analysis such as survey of literature and review of the student assignments is often a time a consuming and cognitively demanding task for students, researchers and instructors alike. Content analysis helps to answer the questions such as: What are the main themes and how are they related? Also, why in spite having all the parts an essay does not seem to flow? This prompts an overarching question: What can be done to make the review of written content more efficient?

Academic writing is structured. In general, the problem that the researcher is trying to address is stated at the beginning of the paper, followed by what is already known—the related work. The solution is introduced approximately half way into the paper, and compared to the related work towards the end—the discussion section. This suggests that different themes are introduced at different times and some themes may overlap.

The process of writing is sequential. A chunk of text does not appear spontaneously, but is formed gradually by connecting words into coherent and rule-guided structures. Writing is like threading beads on a string and if placed on a time line, each word represents a point in time or, using an earlier metaphor, a bead on a string. Since the direction of writing is known, it is possible to identify positions of words, themes, notions and concepts relative to each other or parts of the document. Some words, notions and concepts are auxiliary, whereas some are key to the argument and are carried throughout the text. By identifying tokens (that represent themes) and their positions in text, it is possible to thematically separate textual data into sections, identify main themes and delineate related concepts. Our visualization method termed ThemeTrack allows the user to create a visual map of the textual data and obtain a succinct summary of the information it carries. It also allows the user to get a snapshot of the document without reading it in its entirety, and has several pragmatic implications for the process of learning and teaching: First, it enables researchers, instructors and editors to create visual content summaries. Second, the method may serve as a visual aid for language teaching. Third, the method can be used to make inferences about the document composition.

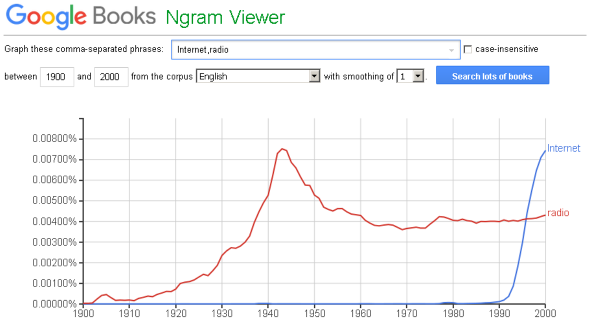
The rest of the paper is organized as follows. In Section 2 we overview the related literature. In Section 3 we introduce the methodology and the algorithm. In Section 4, we discuss the results of the preliminary experiments using a corpus of the real-world texts: graduate level theses and published journal articles, and in Section 5 we conclude with a discussion of future directions.

II. BACKGROUND

Text can be visually represented in a variety of formats. Visualizations provide a relief from the monotony of text and engage the reader to examine the content through a different lens. Visual representation of textual data can be classified into three categories: quantitative, contextual and semantic. The quantitative approach represents text as term counts. Terms could be defined as individual words, or co-locations such as bigrams and trigrams. A common textual visualization technique that uses term counts is the word-cloud (Figure 1). Although various variations exist, the idea behind the word-cloud is that the term frequency determines the term visual properties such as the font size or the color. Function words are often removed as they are frequent and noisy.

Figure 1. Word Cloud visualization of this article.

This approach has been applied in the academic setting and integrated with informal assessment [1], providing the means to visualize and compare students' understanding of course content. One may critique the utility of word-clouds on the basis of their inherent limitation to deliver only a shallow representation of the textual data. Textual data can also be plotted as a time series graph to depict the relationship between the time and token frequency of occurrence. For example Google Ngram Viewer, is a visual information retrieval interface to a corpus of over five million digitized books [2]. It provides a visual representation of the relative frequencies of word collocations in literary works. It is a useful tool for conducting research on social trends as well as linguistic research [3]. Figure 2 depicts the frequency of use of terms: "radio" and "internet" over a period from 1900 to 2000. According to the graph, the term "radio" emerged in literary texts in the early 1900s, peaked in the 1940s and went into decline thereafter. The term "internet" became a subject of growing attention in the early 1990s and the term use has continued to rapidly increase during the next decade.

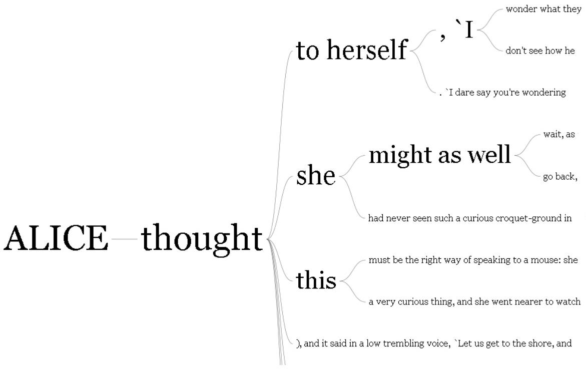


Graph these comma-separated phrases: Internet,radio ☐ case-insensitive

between 1900 and 2000 from the corpus English with smoothing of 1.

Figure 2. Google Ngram viewer.

Terms can also be visualized in a context, relative to other terms. In contrast to the word-clouds, word-trees provide the means to examine the term relationships [4]. Figure 3, depicts words and phrases that follow a root term. This approach has been found useful in literary analysis. Similar to word-trees, tag-clouds depict relationships between the terms [5]. There are different variations of the tree structures offering different functionality. For example, double-trees visualize terms in context as two-sided trees, and are used in linguistic research [6]. They can include the term frequency information for both the words in context and the branching factor.



ALICE
thought
to herself
, `I
wonder what they
don't see how he
. `I dare say you're wondering
she
might as well
wait, as
go back,
had never seen such a curious croquet-ground in
this
must be the right way of speaking to a mouse: she
a very curious thing, and she went nearer to watch
), and it said in a low trembling voice, `Let us get to the shore, and

Figure 3. Word Tree visualization.

This method of visualization is employed in corpus linguistics allowing to efficiently delineate differences between texts [7].

The third type of visualization approach is based on the semantic representation of text. Unlike the previously discussed approaches that organize textual data based on the frequency of occurrence or collocation, semantic-based visualization organizes terms and their relationships based on their meaning. For example, Directed-graphs [8] depict semantic structures by extracting subject–verb–object triplets from each sentence and attaching WordNet [9] synsets (related terms). This is attained through POS parsing and extracting named entities. DocuBurst [10] is a radial graph that depicts an IS-A relationship of a term by utilizing the noun-verb hierarchies of WordNet and term frequencies. A DocuBurst visualization of Leo Tolstoy's War and Peace with war at the root is depicted in Figure 4.

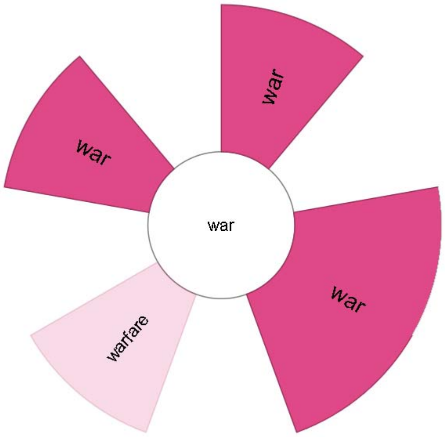


Figure 4. DocuBurst visualization.

III. Implementation

In this section we discuss ThemeTrack, the method for thematic and structural visualization of textual data by mapping cumulative token counts to their relative position in text. In contrast to the existing approaches, ThemeTrack depicts the relationships among user defined terms: their emergence, co-occurrence and decline. The terms may be expressed as lexical features (e.g. word bigrams), they may be expressed as syntactic features (e.g. POS), and also as semantic features (e.g. named entities, sentiments). ThemeTrack visualization can be applied on a single document or a corpus of documents by the one or more authors; and also applied to a class of documents sharing some criteria (e.g. author, subject, genre, etc.). The method allows the user to see how a text is written. It can be considered a visual disassembler that depicts (researcher defined) components of a text; it depicts how the information is organized and presented. The method allows the user to see, in quantitative terms, how one property of a document is related to another. For example several book volumes of one author could be analyzed to identify recurring themes or the use of literary devices within each book and across the volumes. It may also be adopted to analyze other textual data such as music scores. It attempts to answer the questions: what is in the text, and how is it all put together? This will become more obvious as we proceed and review the examples.

The method is comprised of six steps: (a) pre-process text, (b) extract information, (c) count tokens, (d) identify token positions, (e) create pairwise mappings between cumulative token occurrence and their positions, and (f) plot a correlation between the token position and cumulative token count. The output is the two-dimensional x-y graph showing cumulative term frequency on the Y axis and the term position in text on the X axis. The token position is a distance between two events. It can be defined as a temporal dimension (to see which events occur at the same time or over time) or a spatial dimension (to see how far an event is from any part of the document). The argument follows a sequential order and has a clearly defined beginning and an end. Academic texts are comprised of multiple themes; their relationships can be established at any point in text by measuring co-occurrence.

Figure 5 provides an illustrative example of ThemeTrack visualization. In this example, there are three distinct themes. The main theme is present throughout the document (it starts at the position 0 and continues until the very end of the document) and has been repeated 28 times. Theme # 2 is introduced later in the document, and repeated 6 times and does not reoccur later. Theme # 2 is related to the main theme because they co-occur together. Theme # 3 is introduced towards the end. It is not discussed in context of Theme # 2, because there is no overlap, but is related to the main theme. Its frequency of occurrence is higher than that of Theme # 2, so is the time spent discussing it, and therefore it is more prominent than Theme # 2.

Plotting the cumulative frequency of term occurrence and their position produces a two-dimensional x-y graph. The themes are sorted by frequency and the N most frequent themes are plotted. The number of themes is user-defined. For every unique and new token occurrence, the token count increases by 1. The token position (X axis) can be expressed in terms of text length, term count, or as its percentage. Depending on the scale, it can answer questions such as: After how many words a term is repeated, or at what point certain themes converge.

The authors have made the source code available on GitHub (https://github.com/amigal/themetrack). ThemeTrack is available as a Jupyter notebook. The initial implementation was done in the Python language and all tunable parameters are modified in the source code.

A. Data Processing

The raw content comes in a variety of formats and uses various templates. For example, much of the journal articles are distributed in the portable document format (PDF), so are the theses and dissertations, whereas much of the student assignments are distributed in editable formats such as Word documents. The paper layout of journal articles and academic courses vary in the templates they use as well as the style of the bibliographic references. These differences need to be taken into account when processing the raw text, because repetitive headers, footers and citations will contribute to the noise. Once the text is free from noise, it undergoes the information extraction step whose aim it to delineate the main themes and their positions in text. The natural language processing (NLP) techniques provide just that. This can be performed in a variety of ways: ngram based methods, shallow and deep NLP techniques. For example, tokens may be comprised of single words, ngrams (contiguous words or spanning intervening words), POS tags or syntactic ngrams, semantic clusters, etc. In spite of the variations in the information extraction protocol, the underlying concept of the proposed approach is to depict the token lifecycle in relation to other tokens.

In the following sections, we describe the experiments that employ ngram based methods—consecutive word bigrams and trigrams—which perform better when the text is converted to the lowercase and the function words are removed. This entails an additional step of identifying function words, removing them from the text, breaking the text down to individual words and creating word-pairs and word-triplets.

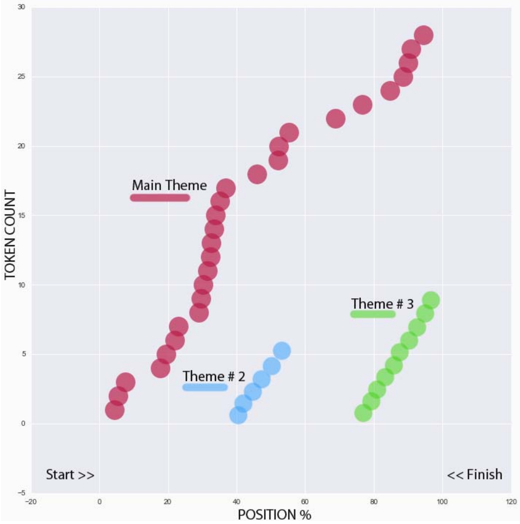


Figure 5. ThemeTrack visualization.

B. Algorithm

The proposed method bears similarity to the frequency based approaches in that the themes are quantified and the magnitude is depicted on a graph. It also bears similarity to the contextual approaches in that the graph depicts the high-level relationships between the themes. It paints a picture of what themes are co-occurring at any point in the text. The proposed method creates a view into the token lifecycle—emergence, reoccurrence and decline.

The method can be algorithmically expressed as follows:

TABLE I. THEMETRACK ALGORITHM

#	Table Column Head
1	Open document
2	Parse document into plain text
3	Remove noise

4	Select information extraction method
5	Tokenize text
6	Create frequency-position matrix
7	Sort by most frequent token
8	**For** each token plot cumulative token count and its position **end For**
9	Close document

IV. Visualizing Academic Content

We have conducted an exploratory analysis using a corpus of the real-world academic texts composed of 20 student thesis and 20 published journal articles. The texts were between 3,000 and 85,000 words. The five most frequent terms were plotted. The thesis dataset was comprised of 10 doctoral-level theses obtained from The Digital Archive of Research Theses of the Open University UK and 10 theses at the master's level, obtained from The Digital Theses library of Athabasca University. The second dataset was comprised of 20 journal articles obtained from the IEEE Xplore Digital Library covering a variety of computer science topics and from The International Review of Research in Open and Distributed Learning covering research in distance education.

Computational and graphing tasks were performed using the Python programming language: using standard libraries for parsing documents into plaintext, regular expressions for filtering out the noise, and the NLTK library [11] for the natural language processing tasks were employed.

The aim of the experiments was twofold. First, to assess the viability of the proposed visualization technique using a dataset of the real-word academic texts. Second, to compare visual representation of information extracted using contiguous bigrams and trigrams to represent themes, and syntactic parsing technique utilizing verb-noun pairs to represent actions. The identified themes and actions were compared against the titles and keywords (for journal articles), which were removed during the pre processing, as were the bibliographies.

A. *Results and Discussion*

Both methods yielded visually similar results, although contiguous word ngrams (N=2,3) captured more meaningful information in the sense that actions (operationally defined as verb-noun pairs) were capturing a lot of abstract information (e.g., giving rise) and therefore insufficient by themselves to yield any concrete inferences about the nature of the text. Future work will address the issue of ambiguity and employ semantic clustering.

The extracted themes were in-line with the keywords and titles. Much of the journal articles exhibited a consistent pattern with one dominant theme carried throughout the document, and supporting themes that emerged and declined, while much of the theses had multiple related themes carried throughout the manuscript. In other words, journal articles exhibited a more sporadic flow of ideas. A sample visualization of a doctoral thesis titled "How does the use of mobile phones by 16-24 year old socially excluded women affect their capabilities?" [12] is depicted in Figure 6.

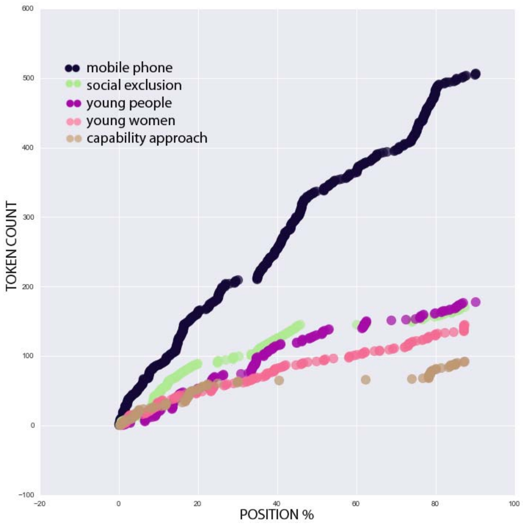


Figure 6. ThemeTrack visualization of a doctoral thesis.

A sample visualization of a journal article entitled "Effective pattern discovery for text mining" [13] is depicted in Figure 7.

These visualizations provide a succinct summary of how content is structured. In the journal article, the terms "text mining", "discovered patterns" and "pattern mining" are semantically similar and occur throughout the document, whereas the term "closed patterns" is introduced at the beginning (around the literature review section) and later reintroduced at the end of the document (around the discussion section). The authors are making a connection between what is new, and what is already known. The notion of "positive documents" is central to this article, it starts after the introduction and merges with the theme of "pattern mining" and "text mining". The size of the journal article is smaller than the thesis, therefore the data points look sparse.

In contrast to the journal article, the sample thesis has a different organizational structure. In Figure 6, we see that all five themes continue throughout the document, with "mobile phone" being the most frequent term. The term "social exclusion" is the second most frequent term at the beginning, and at around half-way through the document, the term "young people" takes its place. The terms "young people" and "social exclusion" show an overlap that is they are both used with almost the same frequency. The term "capability approach" is introduced in the beginning, there are a few instances where it is referenced in the middle of the document, and then reintroduced towards the discussion and the conclusion. During the experiments, there were instances where a new theme would emerge at the very end of the paper, suggesting that the conclusion may not be sufficiently supported, and a revision may be required to further strengthen the argument.

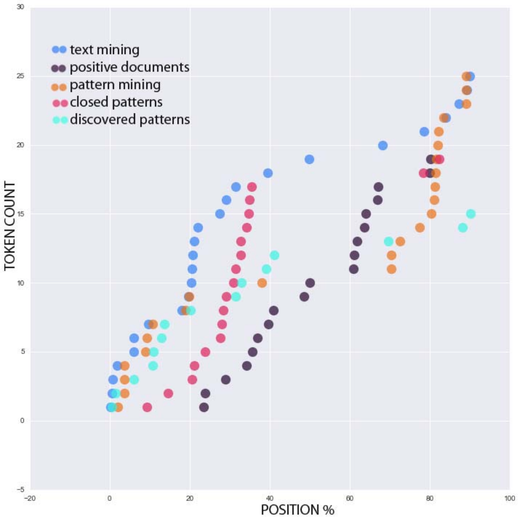


Figure 7. ThemeTrack visualization of a journal article.

Activities such as academic assessments and editorial reviews—that deal with vast amounts of textual data and are concerned with the format or quality of content—may benefit from this approach. ThemeTrack may also serve as a teaching aid, providing instructors with the means of examining thematic arrangement of the student-generated content. This tool may also be of interest to the linguists, literary researchers, psychologists or researchers in any other field that deal with patterns in written narratives and spoken transcripts. Authors have personal preferences for presenting ideas and that leaves a peculiar footprint in textual data that can serve as the basis for authorial-class discrimination. The method may be used to identify intra-authorial preferences in generating content. Further research is required to investigate these issues.

V. Conclusion

In this paper we introduced a novel visualization technique that depicts information on a continuum allowing the user to make inferences about themes and their relationships. The method can be applied to a range of textual data such as movie scripts, song lyrics, legal contracts and musical scores. We conducted a series of experiments using on two datasets comprised of 20 student theses and 20 journal articles. The experiments suggest that the ngram based approach delineates themes more succinctly than the syntactic parsing method using verb-noun pairs. The proposed visualization technique may serve as a teaching aid, or be used to streamline the content review process. The quality of visualization depends on the quality of extracted information. The framework is flexible enough to adopt a variety of information retrieval methods and data types, and future research may compare different approaches and the results they yield. One approach would be to group themes into semantic clusters where similar terms are treated as joint units. Another approach would be to juxtapose information and add sentiment analysis to the thematic flow to track the emotional states in written communication or spoken transcripts. It will also be interesting to examine thematic arrangement as an authorial discriminator in authorship attribution tasks as well as reading comprehension tasks. Our future experimental work will compare the differences of accepted and rejected articles submitted for the peer review journals and also examine the inter-authorial preferences in presenting themes across documents.

Acknowledgment

This work was partly funded by the Spanish Government through the Enhancing ICT education through Formative assessment, Learning Analytics and Gamification project (grant TIN2013-45303-P) and by the Spanish Ministry of Economy and Competitiveness (grant TRA2013-48180-C3-3-P).

References

[1] M. B. Kitchens, “Word Clouds: An Informal Assessment of Student Learning,” College Teaching, vol. 62, no. 3, pp. 113–114, Jul. 2014.

[2] J.B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett,D. Hoiberg, D. Clancy, P. Norvig, J. Orwant et al., “Quantitative analysis of culture using millions of digitized books,” science, vol. 331, no. 6014, pp. 176–182, 2011.

[3] Y. Lin, J.B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov, “Syntactic annotations for the google books ngram corpus,” in Proceedings of the ACL 2012 system demonstrations. Association for Computational Linguistics, 2012, pp. 169–174.

[4] M. Wattenberg and F. B. Vi′ egas, “The word tree, an interactive visual concordance,” IEEE transactions on visualization and computer graphics, vol. 14, no. 6, pp. 1221–1228, 2008.

[5] R. Vuillemot, T. Clement, C. Plaisant, and A. Kumar, “What’s being said near martha? exploring name entities in literary text collections,” in Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on. IEEE, 2009, pp. 107–114.

[6] C. Culy and V. Lyding, “Double tree: an advanced kwic visualization for expert users,” in 2010 14th International Conference Information Visualisation. IEEE, 2010, pp. 98–103.

[7] C. Magnusson and H. Vanharanta, “Visualizing sequences of texts using collocational networks,” in International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer, 2003, pp. 276–283.

[8] D. Rusu, B. Fortuna, D. Mladenic, M. Grobelnik, and R. Sipoˇs, “Document visualization based on semantic graphs,” in 2009 13th International Conference Information Visualisation. IEEE, 2009, pp. 292–297.

[9] G. A. Miller, “Wordnet: a lexical database for english,” Communications of the ACM, vol. 38, no. 11, pp. 39–41, 1995.

[10] C. Collins, S. Carpendale, and G. Penn, “Docuburst: Visualizing document content using language structure,” in Computer graphics forum, vol. 28, no. 3. Wiley Online Library, 2009, pp. 1039–1046.

[11] S. Bird, “Nltk: the natural language toolkit,” in Proceedings of the COLING/ACL on Interactive presentation sessions. Association for Computational Linguistics, 2006, pp. 69–72.

[12] R. Faith, “How does the use of mobile phones by 16-24 year old socially excluded women affect their capabilities?” Ph.D. dissertation, The Open University, 2016.

[13] N. Zhong, Y. Li, and S.-T. Wu, “Effective pattern discovery for text mining,” IEEE transactions on knowledge and data engineering, vol. 24, no. 1, pp. 30–44, 2012.