Auto-Highlighter: Identifying Salient Sentences in Text

Jessica Zeitz Self, Rebecca Zeitz, Chris North
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
{jzself, razeitz, north}@cs.vt.edu

Alan L. Breitler
Simulation and Systems Division
DDL OMNI Engineering LLC
Virginia Beach, VA, USA
alan.breitler@ddlomni.com

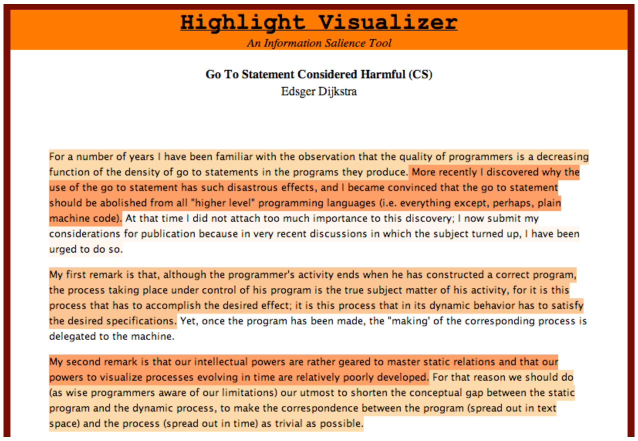
***Abstract*—To help analysts sift through large numbers of documents, we suggest an auto-highlighting system that computationally identifies the topmost salient sentences in each document as a form of summary and rapid comprehension aid. We conducted a user study to gather data about the types of sentences people highlight when reading and comprehending text. Our study focuses not only on the comparison between expert and non-expert users for different document types, but also the comparison between users and common algorithmic metrics for sentence selection. We analyze user-defined categories for describing the variations in the types of highlighted sentences as well as insight concerning rhetoric and language that could strengthen future algorithms.**

Keywords*—*text extraction summarization; user study

I. Introduction

With the rise of big data comes the increasing need for methods to help people rapidly comprehend it. One form this takes is textual data, which often masks its wealth of information through its form and presents the ever present issue of comprehension. Document comprehension is often a challenging task, one that has proven so difficult that humans employ an array of techniques such as annotating, summarizing, and rereading in order to improve understanding. Analysts must routinely sift through numerous documents and determine if and how they are relevant. Given so many documents, an analyst does not have time to read each one in detail. She has to scan at a high level to decide what is worth her time to read. Analysts need new tools that reduce the amount of data to be reviewed and provide an overview to more efficiently assess the documents. Imagine automatically highlighting important information in a document or reducing it to an automatically constucted one-paragraph summary.

We suggest an automated approach that selects the topmost salient sentences from a document. Entire sentences retain the properties of natural text, allowing the analyst to easily read, comprehend and evaluate its possible importance. These salient sentences can be used in or out of the context of the document. In context, the sentences can be visually highlighted, providing cues to important points and aiding in comprehension. By combining these sentences, this provides the analyst with a standalone summary paragraph. These two methods can be used together in user interfaces for document analytics systems, allowing the analyst to interact by progressively increasing or reducing the amount of visible sentences. The original summary sentences provide valuable anchors within the document from which the surrounding context naturally extends by exposing additional sentences or the full text.



Highlight Visualizer
An Information Salience Tool

Go To Statement Considered Harmful (CS)
Edsger Dijkstra

For a number of years I have been familiar with the observation that the quality of programmers is a decreasing function of the density of go to statements in the programs they produce. More recently I discovered why the use of the go to statement has such disastrous effects, and I became convinced that the go to statement should be abolished from all "higher level" programming languages (i.e. everything except, perhaps, plain machine code). At that time I did not attach too much importance to this discovery; I now submit my considerations for publication because in very recent discussions in which the subject turned up, I have been urged to do so.

My first remark is that, although the programmer's activity ends when he has constructed a correct program, the process taking place under control of his program is the true subject matter of his activity, for it is this process that has to accomplish the desired effect; it is this process that in its dynamic behavior has to satisfy the desired specifications. Yet, once the program has been made, the "making' of the corresponding process is delegated to the machine.

My second remark is that our intellectual powers are rather geared to master static relations and that our powers to visualize processes evolving in time are relatively poorly developed. For that reason we should do (as wise programmers aware of our limitations) our utmost to shorten the conceptual gap between the static program and the dynamic process, to make the correspondence between the program (spread out in text space) and the process (spread out in time) as trivial as possible.

Fig. 1. Highlight Visualizer is displaying the first three paragraphs of the technical document, "Go To Statement Considered Harmful" [3]. The highlights depict the important sentences selected by multiple readers; the darker the orange the more readers selected the sentence. In this case, the readers are technical experts.

Previous work has found that intelligent summarizers do outperform summarizers that randomly select sentences [1], [2]. To inform the design of algorithms for automatic highlighting, we look to human behavior. We conducted a user study to discover metrics behind the complex cognitive process humans go through when highlighting and summarizing a document. We aimed to compare user-generated data against algorithm-generated data for extraction-based document summaries. We designed a user interface that highlights sentences based on their salience scores. The interface provides a visual representation of the importance of each sentence in relation to the document, as seen in Fig. 1.

II. Experiment

We conducted an experiment to learn about how and why users highlight sentences in documents, and how that compares to algorithmic approaches. The study focused on two independent variables, *document type* and *participants' area of expertise*, which guided our design. We used one technical (computer science related) article ("Go To Statement Considered Harmful" [3]) and one general or non-technical article ("Time Wars" [4]). We had 40 participants; 20 computer science majors and 20 non-computer science majors. Computer Science participants were considered experts for only the technical document, whereas no participant was considered an expert for the non-technical document.

Participants read each document and highlighted the five most important sentences that would best help them summarize the document. They then summarized each document. We designed the study so that the highlighting was embedded in a larger comprehension task.

III. Findings

As instructed, most participants highlighted 5 sentences. Only 7 and 11 participants highlighted more or less than 5 in the technical and non-technical documents respectively. The technical document contains a total of 47 sentences, of which 25 sentences were highlighted by at least one non-technical participant and 26 sentences were highlighted by at least one technical participant. The non-technical document contains a total of 80 sentences, of which 48 sentences were highlighted by at least one non-technical participant and 52 sentences highlighted by at least one technical participant. Overall, there were 34 sentences highlighted by any participant in the technical document and 66 sentences highlighted by any participant in the non-technical document. As seen in Fig. 2, the highlighted sentences form a power-law distribution such that there are only a few frequently highlighted sentences and most other sentences were highlighted by only 1 or 2 participants. This is true for both documents.

Characterization of Summaries

We sought to answer whether human highlighted sentences are representative of human synthesized summaries and how much readers rely on sentences they highlighted. Elements within each written summary can be traced back directly to the reader's highlighted sentences. We found that the summaries were on average 50% characterized by participants' top five highlighted sentences. The other 50% stemmed from either unhighlighted sentences in the document or synthesized information. Ideas within the summary that could not be traced back to a specific sentence we consider to be concepts synthesized by the participant when reading the document. We conclude that sentences highlighted by a reader are well-representative of the reader's user-synthesized summary. This is solid evidence that selecting salient sentences provides effective document summarization.

User-defined Categories

As people read, certain elements stand out as being more important than other elements. But what characterizes these elements and are these characteristics generally consistent from one person to the next? To answer such questions, we asked participants to explain the reason behind their selection of highlighted sentences. *The analysis helped answer which sentences in a document humans deem salient and why. This will give a basis for defining sentence salience.*

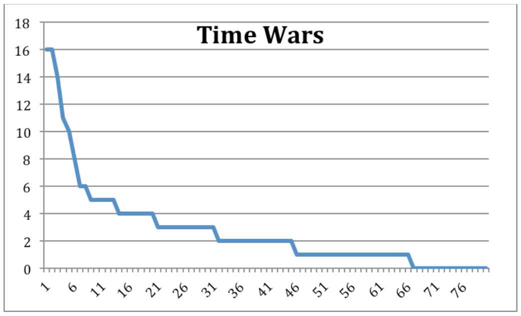


Fig. 2. Power-law distribution for the non-technical document [4], ranking of sentences (X-axis) based on number of participant highlights (Y-axis).

Through open coding, we condensed the reasons for highlighting into 12 categories for the technical document and 21 categories for the non-technical document. These categories were user-defined and stemmed from the reasons participants gave for highlighting certain sentences. Such categories included: introduction, background information, concept connection, example, paragraph summary, and conclusion. The top two categories for both documents were an *argument/main point* sentence and a *supporting evidence* sentence.

Most highlighted sentences were labeled differently by individual participants. Overlaps occurred across the *argument/main point, supporting evidence,* and *conclusion* categories. Even though participants were not completely consistent with reasoning, the sentences were important enough to be highlighted. The *solution* category produced interesting numbers. Seven non-technical participants labeled some sentence in the technical document as a *solution*, however no technical participants used this label. The same is true for the non-technical document; many more non-technical than technical participants categorized sentences as *solutions*. The *profound statement* and *personally resonated* categories produced opposite results; six non-technical participants and six technical participants respectively. We can speculate that some technical participants related to phrases such as "insomniac overstimulation" and having "the urge to check emails" more than others given their involvement in a technical field, Computer Science.

Rhetorical Structure

Categories stem from the elements of the rhetorical structure of a document. We found that categories chosen by participants strongly correlated with elements such as introduction, main point, supporting evidence, and conclusion. These elements are the focus of readers and writers since they provide a basic structure for organization of a document. Categories based on more formalized rhetorical elements (i.e. main point, supporting evidence) were used more often than other elements when labeling sentences. This finding suggests that sentences fitting in one of these main rhetorical elements are more likely to be selected by a reader as salient.

Sentences within these categories also fit rhetorical structure as it pertains to ordering, a phenomenon that occurred in both documents. We found that sentences selected as *introduction* or *background* sentences most often appeared in the first few paragraphs of a document whereas sentences categorized as *conclusion* appeared toward the end.

Experts versus Non-experts

To investigate the imporance of domain knowledge to salience selection, we compared the differences between experts and non-experts when highlighting and summarizing the technical document. Variations between experts and non-experts were minimal. In general, experts and non-experts highlighted similar sentences. The correlation between the two was 0.82; experts and non-experts followed the same overall trend in terms of sentence selection (see Fig. 3).

IV. User and Algorithm Comparison

To mathematically characterize the human highlighted sentences and determine how closely simple algorithm heuristics can mimic the human selection of salient sentences, we compared the user highlights to algorithmically computed sentence salience scores. We computed salience scores for sentences based on several simple and common bag-of-words textual metrics (excluding stop words). There is indication from experimentation in the literature that simple text metrics may be adequate, and in some cases outperform more advanced metrics [5]. We compared these sentence metrics to the count of the number of users who highlighted each sentence (e.g. Fig. 2).

We tested several metrics that attempt to score sentences based on how representative they are of the entire document. The correlation values with the human scores are not high, however the results show promise that with the correct weighted metrics, an algorithm can supply a user with a representative extraction-based document summary. These metrics performed better for the non-technical document. The best-performing metric computed the number of n-grams each sentence shared with all other sentences, with a correlation of 0.61 to human scores on sentences in the non-technical document. A metric which attempted to localize this relevance measure to a window of surrounding sentences performed much worse (correlation of 0.33). The best metric correctly identified about half of the top-most salient sentences as scored by humans, 5/10 and 6/13 top-most salient sentences for the technical and non-technical documents respectively, and correctly eliminated many of the non-salient sentences.

We found that sentence length is somewhat correlated (0.44 for the non-technical document, 0.29 for the technical document) with whether users highlighted the sentence. Users tended to highlight longer sentences. We also tested a metric similar to *tf-idf* called "term frequency / inverse sentence frequency" which measured the uniqueness of each sentence by down-weighting terms that occurred frequently in other sentences. When correlated with the human scores the r value was very low and negative, indicating that users tended to not pick unique or unusual sentences. This along with the results of the other metrics help us understand what "salience" means to users. They tend to highlight sentences that are more representative of the document as a whole instead of sentences that are unique.

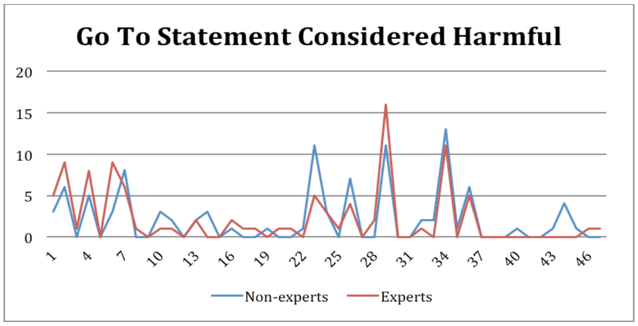


Fig. 3. Overall trend of sentence selection for the technical document. Sentences are in document order along the x-axis and the number of participants who highlighted each sentence is shown on the y-axis.

We conclude that a simple algorithm can automatically highlight sentences to adequately present a summary of a document. Since humans utilize outside knowledge, algorithms cannot exactly replicate sentences selected by a reader. However, incorporating our other findings with these metrics will improve and strengthen extraction-based summarization algorithms to more closely mimic human selection.

V. Conclusions

This study provides valuable information about human-selected salient sentences, human-generated summaries, and the relationships between the two. Humans exploit rhetorical structure to pinpoint salient sentences and then formulate a summary using them. Both experts and non-experts of a document employ these methods. Algorithms using simple text metrics can do a fair job mimicking this human selection. The output is natural to comprehend given it is comprised of complete sentences that can be used as a short summary or as visual highlights in the context of the full text. The findings indicate that such extraction-based summaries composed of salient sentences are well-representative of abstraction-based human-synthesized summaries.

To further develop algorithm effectiveness, we suggest augmenting these techniques with strategies similar to those used by humans. Future work concerning how additional rhetoric and language concepts can improve algorithmic support for comprehension will lead us closer to effectively managing big data.

Acknowledgment

The authors would like to acknowledge and thank Paul Abney and John Rice for their support during this research. This research was sponsored by a Small Business Innovation Research (SBIR)(Phase 1) project funded by the Office of the Secretary of Defense and administered by the Naval Research Laboratory (OSD11-TD1).

References

[1] G. J. Rath, A. Resnick, and T. . Savage, "The Formation of Abstracts By the Selection of Sentences Part I. Sentence Selection By Men and Machines," *Journal of the American Society for Information Science and Technology*, vol. 12, no. 2, pp. 139–141, 1961.

[2] D. Radev, S. Teufel, and H. Saggion, "Evaluation challenges in large-scale document summarization," *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, vol. 1, pp. 375–382, 2003.

[3] E. W. Dijkstra, "Letters to the editor: go to statement considered harmful," *Communications of the ACM*, vol. 11, no. 3, pp. 147–148, Mar. 1968.

[4] M. Fisher, "Time Wars," *Gonzo (circus)*, no. 110, 2012.

[5] X. Wang and A. Kabán, "Model-based estimation of word saliency in text," *Discovery Science*, 2006.