What’s up, Doc? Visualizing information in text documents for better readability and understanding

Nikhil Kini  
University of California, Santa Cruz  
Email: nkini@ucsc.edu

Abstract—The summary and key ideas of a technical paper is textual - in the form of keywords, abstracts, introductions and conclusions. More can certainly be done to improve the presentation of information in long papers with long sentences. We present representations of a document at three granularities - the document level, the paragraph level, and the sentence level to aid in the quicker understanding of the document. This, combined with a semantic graph representation, will enhance the readability of a technical paper.

Keywords—Text visualization, Natural language processing, Semantic graph

I. MOTIVATION

Reading technical papers has been a challenge for me. The format is not arresting enough to catch my attention for long. Sometimes the sentences are so long, that by the time I reach the end, I’m not sure how it began. I presume this might be the problem for others too. At least for those who are just starting out in the world of literature surveys. The only way I can concentrate when I read a technical paper is by writing notes shorthand, which I later use as a quick summary. I believe a more colourful, sparse, summary representation of text would improve the ability of a reader to concentrate on a technical paper, especially when it often contains new words and phrases that one is not familiar with.

In summary, the motivation of the work in this paper is presenting the content of a technical paper with different levels of detail, less amount of clutter, more aesthetically appealing and yet making the information disseminated more effective.

II. APPROACH

We propose a layered presentation of the technical paper, granular at various levels, with a mechanism to drill down to a desired level. We attempt to convert the text into three levels of information.

At the coarsest level, we have the longest representation of the document, the document itself, summarized in the shortest forms possible - n-grams and keywords (which are already present in the paper). This is the view of the document a user will see before it is actually opened. It can provides the authors’ perspective of what the paper is mainly about (keywords) and an automatically statistical measure of the most repeated words (with stop words omitted) in the text.

The next level of drilling down will consider the paragraphs in the paper. Usually, the first and last sentences (or first two and last two sentences) of the paragraph in a well-written paper are representative and should act as summaries. In an attempt to reduce the amount of text presented on the screen, only the first two and last two lines will be presented, and the rest of the text will be replaced by a clickable area. This area will toggle the visibility of the rest of the paragraph. An automated, algorithmic guess for the topic/phrase summary of the paragraph will also be provided.

The last level at which an attempt to improve readability will be made is at the sentence level. Using a parser, we will colour code different components of a sentence that breaks down how the sentence needs to be read at a phase level.

A graph of summarized knowledge presented in the paper will also be presented to the user in the manner of [1]. Nodes in the graph will represent entities and events, and edges will represent relations between any two. It will be made possible to make a node the center of the reader’s attention and present a summary of facts related to this node. It will also be attempted to make important sequences of events easily presented graphically.

III. RELATED WORK

Visualization of textual information content is an active field. In my brief literature survey, I came across a number of papers that are closely related to the work we intend to undertake. A selection of them are presented here. [1] is a document visualization technique for data analysis based on the semantic graph. [2] provides n-gram overview of unstructured text. [3] integrates available NLP technologies into a system that provides novel forms of semantic network representation and visualization. [4] is a visual analytics system for extracting and visualizing temporal trends, clustering, and reorganization in time-indexed textual datasets. [5] provides interactive, topic-based visual text summarization and analysis. [6] presents a system for visualizing the information contained in the text of a web page.

IV. PROPOSED RESEARCH DIRECTIONS

The research in this project will be geared towards making text visualization more readable. Although it can be argued that text in long sentences or paragraphs can be hard to read, it is equally true that taking words out of the context they are in makes interpretation infeasible. We look for an optimization between maximum information presented, readability, and understandability. It will also be a goal to come up with a new textual visualization that is superior to existing methods.
REFERENCES


