Making Results Fit Into 40 Characters: A Study in Document Rewriting

Johannes Leveling and Gareth J. F. Jones
School of Computing, CNGL
Dublin City University
Dublin, Ireland
{jleveling, gjones}@computing.dcu.ie

ABSTRACT

With the increasing popularity of mobile and hand-held devices, automatic approaches for adapting results to the limited screen size of mobile devices are becoming more important. Traditional approaches for reducing the length of textual results include summarization and snippet extraction. In this study, we investigate document rewriting techniques which retain the meaning and readability of the original text. Evaluations on different document sets show that i) rewriting documents considerably reduces document length and thus, scrolling effort on devices with limited screen size, and ii) the rewritten documents have a higher readability.

Categories and Subject Descriptors

H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing—*Abstracting Methods, Linguistic Processing*

General Terms

Experimentation, Measurement

Keywords

Document rewriting, Summarization, Result presentation

1. INTRODUCTION

With the increasing popularity of handheld devices, making effective use of the limited screen size becomes more important. Showing textual information retrieval (IR) results still predominates on mobile devices with small screens. Documents are often laid out for presentation on desktop displays. Access to these documents from mobile devices can result in a poor or unusable experience, because only a limited amount of material is visible to the user and context and overview are lost due to the limited screen size.

We investigate document rewriting, which comprises adding, deleting, or modifying phrases, words, or characters in text to reduce the number of lines needed to render the text while retaining its meaning and readability. The objective of rewriting is to use the display area on mobile devices more effectively, which can have many benefits: elder people or visually impaired people can use an increased font size for better readability; the decision if a document is actually relevant can be made more easily by showing more or longer document snippets; the user effort in scrolling through results is reduced; the list of results can be read faster if shorter words are used; results are easier to understand when more consistent vocabulary is used; and there is less loss of meaning which can result when a full document is summarized. In this paper, we investigate document rewriting and its effect on document length and readability to adapt result presentation to mobile devices.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

2. RELATED WORK

McKay and Watter [4] discuss four different approaches to migrate data to mobile devices: 1. direct migration (i.e. unmodified content), 2. data modification (e.g. summarization), 3. data suppression (i.e. removed content), and 4. data overview (i.e. imposing a hierarchical structure in content to display smaller chunks). Our proposed approach is a mixture of direct migration (keeping unmodified text), data modification, and data suppression. Note that these techniques do not include inserting data (e.g. hyphenation), which is used in the experiments described in this paper.

Traditional summarization and snippet extraction has focused on extracting key sentences from a text [5]. Snippet extraction is a popular method for search engines and question answering systems to show some context for query terms or the exact answers [3]. Most often, snippets are not grammatically correct sentences and include triplets of periods ("...") to indicate sequences of omitted words. Summarization for mobile devices can even limit the result presentation to a single sentence or key phrase [2, 1], which results in the loss or change of meaning.

In contrast, our proposed approach focuses on keeping the meaning and removing only irrelevant or redundant parts of the text in order to shorten it. Document rewriting can even be applied on an already summarized text to shorten it even further. Thus, our goal is not to summarize a text, but to make it fit into a limited display.

A related problem is the reduction of text to make it fit into a given character limit by introducing abbreviations or omitting characters e.g. SMS or Tweets. However, this typically results in text which is more difficult to comprehend (i.e. so-called "textese").

To the best of our knowledge, the application of document rewriting techniques for adaptation of result presentation to mobile devices has not been investigated.

3. DOCUMENT REWRITING

We view document rewriting of text as the insertion, deletion, or modification of characters, words, or phrases of text. (Eliminating full sentences would result in summarization). We investigate the following rewriting methods in our experiments:

Synonyms: replace a word with its shortest synonym from WordNet 3.0[1], generating the correct full word form based on the CMU English morphology database (e.g. "vehicle" →"car").

[1] http://wordnet.princeton.edu/

Acronyms: replace an expanded acronym with the shorter acronym, based on the Vera acronym database[2] (e.g. "United Nations" →"UN").

Simplifications: apply substitution rules compiled from stylistic guidelines and data from the classic Unix tool "diction"[3] (4019 rules), which provides support for simplifying and shortening text (e.g. "in fact" →" ", "red in color" →"red").

Parentheses: delete expressions enclosed in parentheses (e.g. "Smyth (age 43) ..." →"Smyth ...").

Numbers: strip commas from numbers and abbreviate numbers (e.g. "thirteen" →"13", "100,000" →"100K").

Whitespace: remove double or redundant whitespace (e.g. space before a comma).

Hyphens: hyphenate words at end of line, based on TeX hyphenation patterns (e.g. "libraries" →"li-\nbraries").

For our experiments, we assume that a monospaced font is used (e.g. the characters "m" and "i" will take up the same width), with 40-80 characters per line. We think that this is a realistic assumption for phones with limited graphics capabilities and even for modern smartphones, because images or advertisements could take up additional space.

We employ a greedy algorithm for text wrapping, inserting a linebreak when adding a word to the current line would exceed the maximum number of characters per line.

For simplicity, we assume that the first word sense (typically the most frequent one) is correct. We presume that the accuracy of word sense disambiguation is unrelated to the number of lines or the average word length which is used in the computation of readability scores. However, we employ morphologic information to generate the correct word form of the replacement. For example, if the word "added" in the sense of "respond" is encountered in the text, the morphological information includes the base form "add", the part of speech, person, number, gender, case, and tense. The WordNet synonym with the shortest full form matching the morphologic information of the original word (i.e. "say") is then selected as a replacement (i.e. "said").

4. EXPERIMENTS AND RESULTS

We evaluate our proposed approach to document rewriting using two types of metrics: readability scores, which indicate how easy a text is to understand, and the amount of text, which indicates how much user effort is required to scroll through a list of results. We calculate Flesch Reading Ease (FRE, where higher values correspond to better readability), Coleman-Liau grade level (CLGL), and SMOG index[4]. FRE is the standard readability formula and is widely by government agencies, including the US Department of Defense. SMOG and CLGL both estimate the grade level (i.e. years of education) a person needs to understand a text, but CLGL relies on computing the number of characters instead of syllables per word (which can be computed more accurately than the syllable count or sentence length).

In addition, we compute the number of lines (# lines) required to render the text for the original and the rewritten document using 40, 60, and 80 characters per line.

The evaluation is based on three datasets which have been used for summarization evaluations: a sample of texts from the Brown corpus (30 documents), a set of full articles from JAIR (141 documents), and a random set of Wikipedia articles (512 documents). Evaluation results for original (O) and rewritten documents (R) are shown in Table 1. The asterisk (∗) indicates a significant improvement over the corresponding metric for the original text (Wilcoxon test with 95% confidence measure). Note that improvement means a higher FRE score and lower values for the other metrics. In general, rewriting the text reduces the amount of text while increasing readability. The number of lines is reduced by 7.4-13.3% for the tested documents. This can be partially attributed to hyphenation, which reduces unused whitespace before linebreaks. Surprisingly, document rewriting also results in increased readability, because rewriting typically involves removing redundant words and selecting shorter words with fewer characters or syllables. Readability scores such as FRE are based on average word or sentence length, which are decreased by document rewriting.

Table 1: Readability scores for rewriting documents.

Corpus		FRE	CLGL	SMOG	#lines @ 40/60/80 chars		
Brown	O	56.84	11.84	12.62	9495	/6226	/4634
Brown	R	63.83*	10.20*	11.49*	8694*	/5746*	/4291*
Wiki	O	44.93	13.24	13.04	111156	/72804	/54158
Wiki	R	51.79*	11.93*	12.32*	96707*	/63874*	/47726*
JAIR	O	48.59	11.82	13.72	282132	/184682	/137270
JAIR	R	54.34*	10.74*	12.90*	244570*	/161429*	/120513*

5. CONCLUSIONS AND OUTLOOK

We have shown that applying document rewriting can considerably reduce the amount of text (up to 13.3% compared to the original) while improving the readability of documents. This makes document rewriting a viable approach to adapting result presentation to devices with limited display size, which can be applied in addition to summarization to further increase the compression rate.

Future work will include conducting user studies (e.g. measuring reading time and understanding of rewritten documents, when more aggressive document rewriting techniques such as leaving out vowels are applied), integrating more advanced natural language processing (e.g. including NLP for word sense disambiguation, named entity recognition, and coreference analysis), and investigating the effect of document rewriting on other areas of IR such as indexing or summarization.

Acknowledgments

This research is supported by the Science Foundation of Ireland (grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (http://www.cngl.ie/).

6. REFERENCES

[1] O. Buyukkokten, O. Kaljuvee, H. Garcia-Molina, A. Paepcke, and T. Winograd. Efficient web browsing on handheld devices using page and form summarization. *ACM TOIS*, 20(1):82–115, 2002.

[2] M. Jones, G. Buchanan, and N. Mohd-Nasir. An evaluation of WebTwig - a site outliner for handheld web access. In *HUC'99*, volume 1707 of *LNCS*, pages 343–345. Springer, 1999.

[3] Q. Li, K. S. Candan, and Q. Yan. Extracting relevant snippets for web navigation. In *AAAI'08*, pages 1195–1200. AAAI, 2008.

[4] B. Mackay and C. Watters. The impact of migration of data to small screens on navigation. *IT & Society*, 1(3):90–101, 2003.

[5] M. Sanderson. Accurate user directed summarizatiion from existing tools. In *CIKM 1998*, pages 45–51. ACM, 1998.

[2] ftp://prep.ai.mit.edu/pub/gnu/vera/

[3] http://ftp.gnu.org/gnu/diction/

[4] http://www.readabilityformulas.com/