Web Document Summarization by Context

J.-Y. Delort
LIP6 - UPMC
8 rue du capitaine Scott
75015 Paris, France
jean-yves.delort@lip6.fr

B. Bouchon-Meunier
LIP6 - UPMC
8 rue du capitaine Scott
75015 Paris, France
bernadette.bouchon-meunier@lip6.fr

M. Rifqi
LIP6 - UMPC
8 rue du capitaine Scott
75015 Paris, France
maria.rifqi@lip6.fr

ABSTRACT

This paper adresses the issue of Web document summarization. We consider the context of a Web document by the set of pieces of information extracted from the content of all the documents linked to it. We put forward two new summarization by context algorithms. The first one uses both the content and the context and the second one relies only on the elements of the context. It is shown that summaries based on the context are usually much more relevant than those only made from the content of the target. Optimal conditions on the size of the content and the context of the document to yield the best summaries are studied.

Keywords

Summarization, Web documents, Websites, context.

1. INTRODUCTION

Summaries are a key factor of the current usability of the Web. Applications range from snippet generation by search engines [1], to the tailoring of the content of Web documents to suit specific displays [3] (e.g. PDA) through Web document summarization in an accessibility purpose (e.g. for blind people).

Automatic summarization researches have been mainly focused on plain-text documents. Their usability with Web pages is limited because: Web pages are made up of elements which cannot be summarized (such as pictures), textual information is often scarce, and, though they are human-readable it is hard to make a generic computer program able to distinguish between relevant and shallow information in an HTML document. For a few years, more and more Web applications have successfully taken into account the context of a document instead of the document itself [4, 2]. Here we consider the context of a Web document by the set of pieces of information extracted from the content of all the documents linking to it.

This paper introduces first the main issues of summarization by context. Two summarization by context algorithms are outlined. The first one combines both the content and the context of a document while the second is only based on the context. These algorithms have been compared with a classical content-based algorithm.

2. GENERAL ISSUES

WWW2003, May 20–24, 2003, Budapest, Hungary.
ACM xxx.

Any context-based summarizer has to face three different kinds of issues:

contextualization Extracting the pieces of information among the documents of the context which are dealing with or informative about the target.

partiality Sometime the pieces of information among the documents of the context are only stressing on a part of the content of the target. They must be then put together in a way they cover entirely the target.

topicality The elements of the context have to be distinguished between those that are related to the target but do not contain any clues saying what the target is about and those the content of which gives an overall insight into what the target is dealing with. This difference is illustrated by the following example:

1. $< LINK > CNN < /LINK >$[1] *reported the rate of cars robbed in Nevada has increased of* 0.1 *in the second quarter.*
2. $< LINK > CNN < /LINK >$ *is a news website.*

3. SUMMARIZATION ALGORITHMS

The summarization algorithms presented here are using the sentences extracted from the context of a document after a preprocessing contextualisation step that is not described here. The topicality issue has been formalized as follow: A *reference sentence* defines a sentence the content of which does not contain any clue saying what the target is about (for instance, the first sentence in the CNN example) and a *subject sentence* corresponds to the situation where the content of the sentence gives an overall insight into what the target is dealing with (the second sentence in the previous example). Clearly these definitions are not crisp. Indeed, to what extent will we consider a sentence to give a clear enough representation of the target? This leads us to define the *degree of topicality of a sentence S with a document D* by a number $T(S, D)$ between 0 and 1 such that $T(S, D) = 0$ means that the sentence is a reference to D and $T(S, D) = 1$ means that the sentence is a subject of D. The two proposed algorithms are outlined in the two following subsections.

3.1 Mixed approach

The first algorithm consists in computing, for each sentence of the context, its degree of topicality. To be efficient

[1] links to http://www.cnn.com

this method requires that: 1) the target page can be fetched and contains textual information and, 2) this information is sufficient to represent the content of the document. Rifqi *et al.* [5] have proposed a definition of satisfiability that suits the definition of topicality of a sentence with a document: A measure of satisfiability *"corresponds to a situation in which we consider a reference object or a class and we need to decide if a new object is compatible with it or satisfies it"* [5]. If the content of the target is sufficient it can be considered as the reference object. Satisfiability measures can be used to compute the degree of topicality provided that the documents and sentences are considered as sets of words. The chosen degree of topicality of a sentence S with a document C could be given by the widespread satisfiability measure: $T(S,C) = \frac{|S \cap C|}{|C|}$. The mixed summarization algorithm works as follows:

1. Compute the degree of topicality of each sentence with the target document,
2. Rank the results with respect to these values,
3. Select the sentences having the best topicality values for the summary.

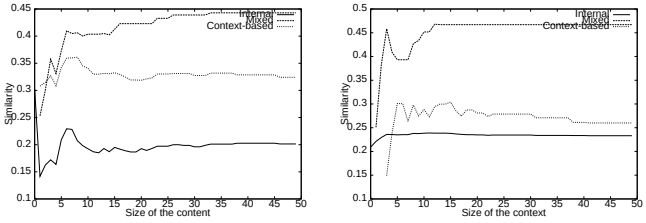
3.2 Context-based approach

When the content of the target document is too scarce the previous method cannot be applied to it - neither can any method using the content of the target as an input. The second method proposed here is based on the following hypothesis: usually, among the sentences of the context of a target document, contents of the subject sentences are closer than those of the reference sentences. In other words, the terms chosen to describe one page cannot be very different (they can yet be synonyms). On the other hand, there are plenty of reasons to quote a website or a Web page without saying what it is about. Thus this method has a clustering step during which sentences are clustered with respect to their content. The context-based summarization algorithm works as follows:

1. Cluster the sentences with respect to their content (we used a hierarchical clustering algorithm to carry out this task for it does not require the number of cluster to be chosen *a priori*),
2. Rank decreasingly all the clusters with respect to the number of sentences they contain,
3. Select within the largest clusters a sentence found by means of an internal ranking function to be kept for the summary.

4. EVALUATION

The classical summarization approach - here refered to as *internal* approach, computes for each sentence of the content a similarity value between the sentence and the whole content. Then the sentences are ranked with respect to their degrees of similarity and those having the highest values are kept in the internal summary. For each page they point to, Web directories also contain a short summary on it (usually no more than one or two sentences). We have tested our different algorithms with a similarity measure, the cosine. The summaries obtained with context-based methods as well as those got with the internal approach are to be compared on the basis of their similarity values with respect to the DMOZ description. The considered testing database contains tuples of three elements *DMOZ summary/content/context* that were gathered thanks to the following process: first, 2000 links with their summaries were randomly taken in the DMOZ repository. Then, the preprocessing contextualization process was applied to each link and the target document download and its sentences extracted (about 80000 documents were thus fetched).



The following table sums up the results previously seen and gives the approach to follow with regard to the size of the context and the content of the target.

		Size of the context	
		≤ 4	> 4
Size of the content	≤ 3	internal	context-based
	> 3	mixed	mixed

5. CONCLUSION

Using context-based approaches seems to be a promising alternative way of Web document summarization. This paper introduced and studied first the main issues of summarization by context. Two approaches were proposed the respective efficiency of which depends on the size of the content and the context of the target document. An evaluation technique derived from intrinsic evaluation techniques of summary has been introduced to face the problem of comparing extracts which sentences come from different sets. Future works will focus on the use of smoother similarity measures and new kinds of representation and on automatical tuning of the parameters of the two methods.

6. REFERENCES

[1] E. Amitay and C. Paris. Automatically Summarising Web Sites - Is There A Way Around It? In *Proc. of the 9th Int'l Conference on Information and Knowledge Management*, 2000.

[2] G. Attardi, M. S. Di, and D. Salvi. Categorisation by Context. *Journal of Universal Computer Science*, 1998.

[3] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices. In *Proc. of the 10th Int'l World Wide Web Conference*, 2001.

[4] J. Furnkranz. Using links for classifying Web-pages. Technical report, Austrian Research Institute for Artificial Intelligence, TR-OEFAI-98-29, 1998.

[5] M. Rifqi, V. Berger, and B. Bouchon-Meunier. Discrimination power of measures of comparison. *Fuzzy Sets and Systems*, 110:189–196, 2000.