Enhanced Web Document Summarization Using Hyperlinks

Jean-Yves Delort, Bernadette Bouchon-Meunier, Maria Rifqi

Enhanced Web Document Summarization Using Hyperlinks

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-UPMC
University Paris 6
8 rue du capitaine Scott
75015 Paris, France
maria.rifqi@lip6.fr

ABSTRACT

This paper addresses the issue of Web document summarization. As textual content of Web documents is often scarce or irrelevant and existing summarization techniques are based on it, many Web pages and websites cannot be suitably summarized. We consider the context of a Web document by the textual content of all the documents linking to it. To summarize a target Web document, a context-based summarizer has to perform a preprocessing task, during which it will be decided which pieces of information in the source documents are relevant to the content of the target. Then a context-based summarizer faces two issues: first, the selected elements may partially deal with the topic of the target, second they may be related to the target and yet not contain any cues about the content of the target.
In this paper we put forward two new summarization by context algorithms. The first one uses both the content and the context of the document and the second one is based only on the elements of the context. It is shown that summaries taking into account the context are usually much more relevant than those made only from the content of the target document. Optimal conditions of the proposed algorithms with respect to the sizes of the content and the context of the document to summarize are studied.

Categories and Subject Descriptors

I.7.m [**Computing Methodologies**]: Document and Text Processing—*Miscellaneous*; I.2.7 [**Computing Methodologies**]: Natural Language Processing—*Text analysis*; H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms

Algorithms, Experimentation

Keywords

Summarization, Web document, Hyperlinks, Context.

1. INTRODUCTION

Summaries are a key factor of the current usability of the Web. Because they are intended to give a quick overview of a document or a Web site, summaries are used to face situations where many documents are involved and need to be either discriminated or, on the contrary, synthesized. Result pages of search engines often include a snippet fitting the user's current interest, for each of the proposed answers. Sub-categories of Web directories containing topic-related URLs also propose a short summary for each of them. Recently, with the development of handheld devices, summarization techniques have been proposed to tailor the content of Web documents to suit very small displays.
Automatic summarization research is more than 50 years old, but, until the development of the Internet, researches were mainly focused on plain-text documents. The interest of existing techniques has been proved very limited with Web pages. A few reasons may explain these disappointing results which are, all of them, related to the content of the document:

- Web pages are multimedia, they are made up of elements which cannot be summarized (such as sound, pictures, video, etc.). Moreover, textual information is often scarce,
- Web pages often deal with very different topics (it would be an interesting challenge to try to summarize the Yahoo homepage) [5],
- Web pages are human-readable but 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 (*i.e.*, related excerpts taken in all the other documents that point to it) instead of the document itself. A few algorithms use both the context and the content [14], others only rely on the context [8, 2].
Here we consider the context of a Web document by the textual content of all the documents linking to it. Summarization by context seems to be a powerful alternative way,

to summarizing by content. It has two interesting advantages:

- The drawbacks and limitations of summarization by content are removed,
- When a document points to another one, it often includes a description of its link to this page. In other words, the context may contain already human-made summaries of the document.

This paper introduces first the main issues of summarization by context. Two summarization by context algorithms are presented. The first one combines both the content and the context of a document while the second takes only into account the context. These new algorithms are compared with a classical content-based algorithm by means of an intrinsic comparison process which is also introduced.

In the sequel a *source page* will be defined, by opposition to a *destination or a target page* T, as a document S that points to T.

2. RELATED WORKS

Summarization by context is concerned with two separated fields, summarization algorithms and context of Web documents. Overviews of existing works for each one are presented. Then, the InCommonSense system which pioneered the issue of summarization by context is detailed. In the last part, the main issues a context-based summarizer has to face are discussed.

2.1 Summarization

Summaries can be distinguished into two categories: abstracts and extracts. An extract is entirely made of text spans extracted from the original document whereas an abstract is a summary, a least some materials of which does not exist in the original document (e.g. point of view on the document, paraphrase, etc.). Abstracting a document requires the ability to manage various hard AI problems such as discourse understanding, natural language processing and abstraction. Thus, most existing summarization algorithms yield extracts instead of abstracts. Roughly, such summarizers assess how relevant each sentence of the document is with respect to a specific or a generic query - which is in this case the whole document. Then, sentences are ranked with respect to their degree of similarity [9]. The main drawback of this method is that extracts are not meaningful because selected excerpts kept in the summary may have no cohesion when put together. As previously explained, such techniques are not suitable for Web documents.
Research on Web document summarization is active [1, 3, 5, 20], even commercial softwares have been recently released[1]. Summarization techniques have been proposed to tailor Web documents to disabled people (e.g. visually impaired people [19]) or to the different kinds of Internet access terminals. Buyokkokten and al. [5] have proposed five approaches fitting the specific constraints of the screens of handheld devices to summarize Web documents. First, they split a document into fragments called semantic textual units (STU). Then each STU is summarized using keyword or sentence extraction or both of them. In particular, they describe an algorithm of sentence selection based on clusters of the most significant words. Their approach has been tested and extended by Zhang [20] who has proposed a system to automatically summarize a complete Web site using the same frame as their sentence selection algorithm but also including other factors such as page length and depth. Berger and Mittal [3] have worked on models automatically generating the "gist" of a Web document. A gist is midway between an extract and an abstract (all the terms come from the documents but the gist has been generated and is not present in the document).

[1] see for instance the paper of Jones [11] for a broad overview of existing summarizers.

2.2 Context of Web documents

With the growing number of applications successfully relying on the context, a few works have tried to understand the reasons of such a success. Attardi [2] has explicitly described the two basic implicit hypotheses of *characterization by context*:

1. If a source document points to a target document, then the context of the link in the former should be connected to the content of the latter,
2. The whole context of a document is sufficient to discriminate it.

Lately, Menczer [12] christened the first one, the link-content hypothesis and the second one, the link-cluster hypothesis and he proposed mathematical definitions to formalize them. Davison's work [7] on topical locality has proved the validity of the first one. Davison demonstrated the discrimination power of the anchor text[2] on the linked document. This was then confirmed in additional papers [6, 12]. Yet, with an average size of 2.7 words [7], anchor texts are not sufficient to make interesting summaries. Furthermore Davison reported that, most of the time, anchor texts are often containing just the title or the URL of the target document.

2.3 The InCommonSense system

With the InCommonSense system, Amitay and Paris [1] pionnered summarization by context. The purpose of their system is to propose snippets for search engine results. The context is first gathered by means of a query of the type "link:URL" to a search engine. Then segments of text containing the link to the destination URL are extracted. Eventually a description filter process chooses the most accurate snippet among those in the context. The authors have led a survey of more than 700 individuals looking to compare the InCommonSense system with AltaVista and Google link summarizers[3]. The aim was to evaluate "how easy it was to find the information needed with regard to the snippet generator used?". The authors reported that, in average, people

[2] anchor text is the text encapsulated in the tags <A> and </A> which are used to link to another document in HTML.

[3] Altavista snippet generator takes the most relevant sentence in the target document. Google take in the target document several text spans around the words of the initial query.

have preferred the InCommonSense system. The system suffers however from its drastic sentence selection process that prevents it from summarizing pages when the context is not very large. Furthermore, the system is intended to make one-sentence sized snippets and tuning the system for generating longer and more detailed summaries is not described. Finally there are two important, specific to summarization by context issues that the system does not handle. We call them partiality and topicality issues and we define them in the next subsection.

3. SPECIFIC ISSUES

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 cues about the content of the target 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* $>$[4] *reported the rate of cars robbed in Nevada has increased of* 5% *in the second quarter.*
2. $< LINK >$*CNN*$<$ */LINK* $>$ *is a news website.*

In the next sections, these issues will be discussed.

4. CONTEXTUALIZATION ISSUE

As previously seen, Davison's works have shown that anchor texts are related to the content of the target but they do not convey enough information on it. On the other hand, the problem with taking a chunk of words around the links is that, at the best, this is very likely to dilute the relevant information on the target, and at the worst, this would produce something meaningless. Accordingly, a compromise between anchor texts and text spans consists in taking the whole sentence containing the link to the target. Sentences, as basic units of natural language, are likely to convey more information than anchor texts and to be easily understandable by human beings.

The contextualization process of a document refers to all the intermediary steps to gather the sentences of its context. A query of the type *link:URL_of_the_target* is used to learn the URLs of the pages pointing to the target from Google. Then, source pages are fetched and text spans surrounding the links to the target, called pseudo-sentences, are extracted. In order to be as objective as possible, all the documents of the context belonging to the same domain (having the same prefix) are removed. A pseudo-sentence will be considered as a sentence provided that it has some specific syntax features. Chosen features will be described in section 7.

Sentences will be represented in the vector-space model [16]. In this model, a sentence is represented by a vector of weighted terms $\overrightarrow{s}$. In the TFIDF representation, a widely used weighting system, a term weight is defined as:

$$w_{ik} = \frac{tf_{ik} . \log(N/n_k)}{\sqrt{\sum_{j=1}^{N} (tf_{ij})^2 . \left(\log(N/n_j)\right)^2}}$$

where tf_{ik} is the frequency of occurrence of the term W_k in sentence S_i ($tf_{ik} = 0$ if W_k is not present in S_i), N is the size of the context and n_k is the number of documents in the context with term W_k. One of the interesting things with this method is its ability to discriminate the sentences: the TFIDF weighting system gives large weights to words that occur frequently in particular documents but rarely among the others. Sometime this property must be avoided, for instance if one is interested in the similarity between two sentences. Then the TF representation can be used instead. The corresponding weighting system for the TF representation is given by: $w_{ik} = tf_{ik} / \sqrt{\sum_{j=1}^{N} (tf_{ij})^2}$.

In the sequel, the TFIDF and the TF representations as well as the boolean one (assigning 1 if the word is present, 0 otherwise) weighting assignments will be used according to the given situation.

5. PARTIALITY ISSUE

The partiality issue can be addressed by extracting the *representants* from the context of a target document. The set of representants of a given context is the smallest subset of sentences of the context such that removing one element from it would make the overall context information decrease. These not-removable sentences will be referred to as *representants*.

Consider the three following sentences taken from the context of the Lance Armstrong Foundation homepage[5]:

1. *Visit the* Lance Armstrong Foundation website,
2. *To find out more about Lance and his amazing career, visit* www.usacycling.org/?upload/armstrong.html *and visit* www.laf.org *to learn about or contribute to the Lance Armstrong Foundation, dedicated to helping people manage and survive cancer.*
3. *The* Lance Armstrong Foundation *helps people survive and manage cancer.*

Note that sentence 2 provides the more comprehensive pieces of information and the contents of sentence 1 and 3 are included in sentence 2 (if one takes into account the synonymy). Removing sentences whose content is included in another sentence of the context is an easy way to reduce its size without loss of information. It can be compared by taking the free elements of a subset of a vectorial set.

To compare the inclusion degree of two sentences let us introduce an inclusion measure I of one sentence in another.

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

[5]http://www.laf.org

For instance given two sentences I_i and I_k, the inclusion value $I(S_i, S_k)$ of I_i in I_k could be defined by:

$$I(S_i, S_k) = \frac{\sum_{j=1}^{N} w_j^i . w_j^k}{\sum_{j=1}^{N} w_j^i} \quad (1)$$

Let $\mathcal{S} = \{S_i\}_{i=1..N}$ be the context of a document. Sentences which can be removed from the context without loss of information are defined by the set:

$$\mathcal{S}' = \{S_i \ : \ \exists k \neq i, \ I(S_i, S_k) = 1\}$$

Then the representant set is $\mathcal{S} - \mathcal{S}'$. First, if two or more sentences have identical representations (their degrees of inclusion are equal), only one will be kept. Then, let us consider the no-symmetric $N' \times N'$ inclusion matrix $M = [M_{ik}]_{i,k}$ with $M_{ik} = I(S_i, S_k)$ (where N' is the number of sentences after removing identical sentences) and associate the following vector with it:

$$\mathbf{X} \quad = \quad \begin{pmatrix} X_{1j} \\ \vdots \\ X_{N'j} \end{pmatrix} \quad (2)$$

where X_{ij} denotes the number of elements in the matrix M at the row i having an inclusion value equal to 1. The representant set consists in all the elements $S_i \in \mathcal{S}$ such that $X_i = 1$.

5.1 TOPICALITY ISSUE

The topicality issue has been formalized as follow: A *reference sentence* defines a sentence the content of which does not contain any cues about the content of the target (for instance, the first sentence in the CNN example). A *subject sentence* corresponds to the situation where the content of the sentence gives a good insight into what the target is dealing with (the second sentence in the CNN 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 presented in the two following subsections.

5.2 Algorithm 1: mixed approach

The first algorithm consists in computing, for each sentence of the context, its degree of topicality. To be efficient 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.* [15] 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"* [15]. 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.

5.3 Algorithm 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 chosen clustering approach belongs to the class of hierarchical clustering algorithms [10]. In our situation, hierarchical clustering is interesting for several reasons: 1) it does not require the number of clusters to be chosen *a priori*, 2) it is easily tunable and thus yields more natural classes. First a (symmetric) similarity function must be chosen: here, the chosen document representation is TFIDF and the similarity measure used is the classical cosine: let S_1 and S_2 be two sentences represented by the vectors $< w_1^i, .., w_N^i >$ and $< w_1^k, .., w_N^k >$ respectively, their similarity value is given by:

$$Sim(S_i, S_k) = \frac{\sum_{j=1}^{N} w_j^i . w_j^k}{\sqrt{\left(\sum_{j=1}^{N} w_j^i\right)^2 . \left(\sum_{j=1}^{N} w_j^k\right)^2}}$$

Our algorithm works as follows for a summary of maximum size l and a context $\mathcal{S} = \{S_i\}_{i=1..N}$ (the first 4 steps are a simple instantiation of the hierarchical clustering algorithm):

1. Assign to each sentence its own cluster and define the similarity between two clusters $\{S_i\}$ and $\{S_k\}$ as the similarity value $Sim(S_i, S_k)$,
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that there is now one cluster less,
3. Compute the similarities between the new cluster and each of the old clusters,

4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N or the similarity value of the most similar pair of sets is smaller than a given threshold $0 \leq \alpha \leq 1$,

5. Remove all the one-element sized clusters,

6. Rank decreasingly all the clusters with respect to the number of sentences they contain, let $\{C_1, .., C_p\}$ be the order set of the remaining clusters,

7. Apply a ranking function f within each cluster C_i,

8. While $i < min(l, p)$ take the element of C_i having the highest value for f.

Step 3 can be done in different ways, for instance one can consider that the similarity between two sets is equal to the highest similarity value of a pair of elements taken in each cluster (this is called simple link distance, SLD), or, on the contrary one can take the smallest similarity value (called complete link distance, CLD), or, one can take the average similarity value that is the average similarity value from any member of one cluster to any member of the other (average link distance, ALD). With steps 7 and 8, the algorithm seeks to gather in the summary sentences that are likely to be complementary. The f function uses the length and the closeness of the sentences to the centroid[6] of the cluster to rank all of its elements. Before computing the value $f(S_i)$, the sentences of the clusters are decreasingly ordered with respect to their size and their closeness to the center of the cluster in two hashtables $R1$ and $R2$ where $Rx[S_i]$ is the rank of the sentence S_i for $x = 1, 2$. We discovered that the value $f(.)$ was generally good when, for each sentence S_i, it was given by $f(S_i) = R1[S_i] \times \sqrt{R2[S_i]}$.

5.4 A comprehensive example

In this section a comprehensive example is presented to illustrate the behavior of the previous algorithms. The Web page considered is the homepage of the Journal of Young Investigator[7]. The purpose of this website is to promote undergraduate researches in science and engineering. The raw content of the homepage contains 59 pseudo-sentences and only 28 sentences. Remaining sentences are titles of articles or news available in the website and they do not explicitly deal with the purpose of the website. Thus a content-based summarization algorithm would not yield relevant summaries for this document. Now let us look at some of the sentences of the representant set. The representants remaining after a contextualization step are ranked in [Fig. 1] with respect to their degree of topicality.

In the DMOZ directory, The Journal of Young Investigators is described as *"An online journal dedicated to the presentation of undergraduate research in science, mathematics, and engineering. JYI provides an opportunity for undergraduates to participate in the entire scientific enterprise."*.
The closeness of the first sentences with this summary in [Fig. 1] shows why the first method is interesting.

[6]The centroid is given by the vector which weights are the average weights of all the words of the representation of all the sentences within the cluster.
[7]accessible at http://www.jyi.org/index.html.

1	Journal of Young Investigators - Online journal dedicated to the presentation of undergraduate research in science, mathematics, and engineering.
2	National Journal of Young Investigators - "an entirely undergraduate effort aimed at showcasing research conducted by undergraduates, building a sense of community among undergraduate scientists, providing services and information useful to those students, and enhancing the contribution of undergraduates to the larger scientific community".
3	National Journal of Young Investigators (JYI) [http://www.jyi.org/] 1998+, full text; "a faculty and student reviewed, peer edited and published, national journal" of science and engineering, freely available.
4	Journal of Young Investigators http://www.jyi.org/ A free, undergraduate, peer-reviewed science journal, "JYI provides opportunities for students to participate in the scientific review and publication processes, primarily through the operation of its peer- reviewed journal for undergraduates."
5	The first was the Journal of Young Investigators (JYI), an online undergraduate research journal that seeks to promote the publication process as an integral portion of a complete science education.
6	Along with this sort of independent work, I joined an undergraduate science journal that William Head first brought to my attention, the Journal of Young Investigators.
7	For more information, or to view our current issue, please visit: http://www.jyi.org/.
8	Contact the National Journal for Young Investigators at http://www.jyi.org.
9	Called the National Journal of Young Investigators (JYI), the Web-based publication premiered in December at http://www.jyi.org.

Figure 1: Context of $http://www.jyi.org/index.html$

With α set to 0.1 the second method proposes only three clusters of size two and all the others sentences are moved away in one-element sized clusters. The 2-element-sized clusters were ordered: $C1 = \{2, 4\}$, $C2 = \{5, 1\}$, $C3 = \{3, 8\}$ after computing the inside-cluster ranking function.

6. HOW TO EVALUATE CONTEXT-BASED SUMMARIZERS?

Existing evaluation techniques of summarization algorithms may be distinguished between two classes [18]. *Extrinsic* ones focus on the efficiency of the system in achieving a specific task, for instance classification. The second kind of evaluation techniques, called *intrinsic*, try to assess the inner qualities of the summaries. They require a testing dataset of document summaries that will be considered as models. Existing intrinsic approaches were mainly designed for the "classical situation" where the models are extracts made by experts and summarization algorithms output extracts too. Under these conditions, recall and precision rates can be computed because the sentences of the models and those of the summaries come from the same document.
There exists a great deal of databases of summaries of websites and Web documents on the Web that could be considered as models. For instance, for each page they point to, Web directories also contain a short summary on it (usually no more than one or two sentences). Still, intrinsic evaluation approaches using recall and precision rates cannot be generalized here with these databases for the algorithms are based on sentence extraction within different sets (the content and the context) and the model is made up of sentences which belong to none of these sets. Thus, we propose to use a similarity function to remove this problem. This similarity value will be taken instead of the recall and precision rates and computed with respect to the model.

7. EXPERIMENTATION

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 contextualization process was applied to each link and the target document download and its sentences extracted. About 80000 documents were thus fetched.

7.1 Preliminary steps

We developed a wrapper to carry out the contextualization step for the target documents. First, dynamic content of the source documents was removed. Then, as in [5], pages were splitted into fragments of text using structural information (bullets, paragraph, section...). These fragments comes down to the STU seen in subsection 2.1. Abbreviations and numbers containing dots were then detected and encoded without punctuation. Then, each STU was splitted into pseudo-sentences. Pseudo-sentences containing more than 7 links, or with a total number of words larger than 50 or less than 3 will be removed for it is very unlikely they are sentences.
Once all pseudo-sentences were collected, a syntax filter was applied to them. This filter uses a part-of-speech tagger, called TreeTagger [17], to annotate each word of the pseudo-sentences. TreeTagger is able to recognize proper nouns and has a very comprehensive English language dictionary. Any word not included in the dictionary will be tagged <*unknown*>. A pseudo-sentence was considered as a sentence provided that it contains a verb and that the number of unknown tags does not exceed 4.
Only words tagged as adjectives, verbs or nouns by the part-of-speech tagger have been be kept in the term-space provided that they were not previously stopped by a stopword list. Eventually the final representations were expanded *via* the synonym thesaurus WordNet [13] and normalized.
To compare the inclusion degree of two sentences, the boolean weighting system was chosen and the inclusion measure is the one given by equation 1.

After the contextualization process about 80% of the elements were removed from the retrieved contexts. This means that documents are mainly linked with a picture or a very short description but not with a sentence. After the representant filtering step, the average context size was: 5.9.

7.2 The internal approach

Context-based algorithms were compared with a classical content-based algorithm. Content-based algorithms - here we will refer to them as the *internal* approach, compute 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. As this is usually done with this approach, the representation chosen here is TFIDF and the similarity measure, the cosine.

7.3 Results and discussion

The summaries obtained with the 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 summaries. To compute the similarity between two summaries, the TF representation will be used and the similarity measure is the cosine. In the following experimentations, the minimum degree of topicality for a sentence to be maintained in a summary is set to 0.1.

Figure 2 shows the similarity values when the minimum number of sentences in the content of the target changes. The default size of the summaries was chosen equal to 2 and the document tested have a context size larger than 4. The first method and the internal one have their similarity values improved increasingly. However, internal approach is getting better more slightly than in the "classical situation" where usually the slope is deeper. The average similarity values of the second method is the same when the content is larger than 1 or larger than 30 - what was foreseeable for this method is independent of the content. Another reading that can be made from this chart is that, when the number of sentences in the content is larger than 3 the first method should be preferred to the two others, otherwise the second method yields the best summaries.

Figure 3 shows the similarity values when the minimum number of sentences in the context of the target changes. The default size of the summaries was chosen equal to 2 and the tested document have a content size larger than 4. Under these conditions, the first method is better than the two others. When the size of the context is equal or larger

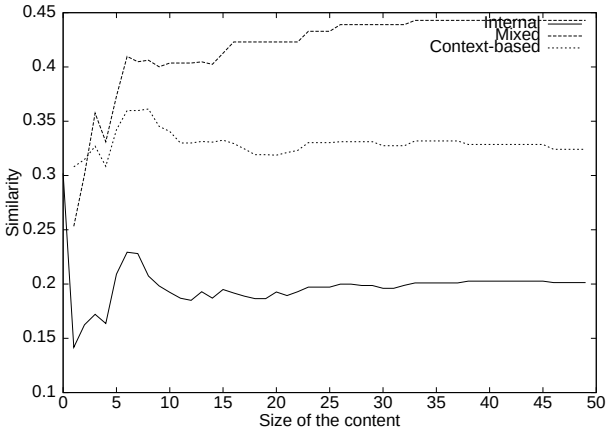

Figure 2: Sim. values when the content size changes

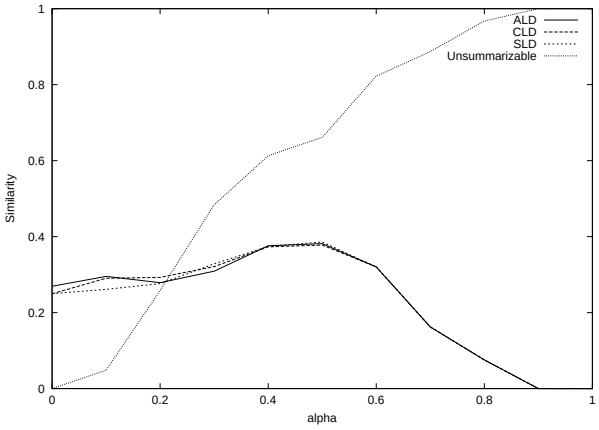

Figure 4: Second method with different parameters

than 3 the first method is the best and the second method beats the internal one when the size of the context is larger than 4. The first method must be preferred when the sizes of both the context and the content are high. When the context is almost empty and the content is not, none of the first and second methods can be used and the internal approach is the only one possible. When the content is almost empty but the context size is large it is better to use the second method.

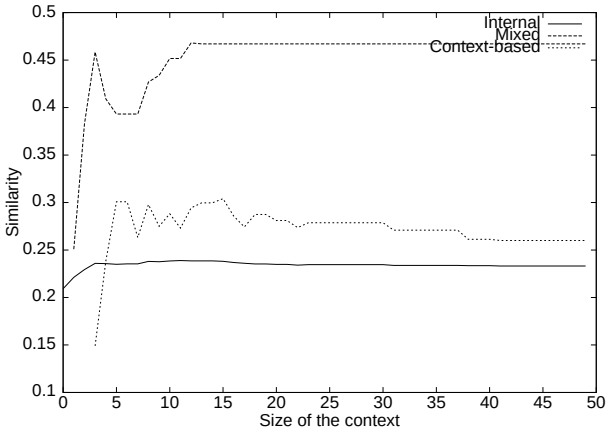

Figure 3: Sim. values when the context size changes

Figure 4 compares the similarity values obtained with the second method when the method for computing the distance between the clusters (used at step 3 in the algorithm) and the parameter α change. We see that whatever the chosen method is, the similarity remains almost unchanged. The curve called "Unsummarizable" gives the proportion of situations when the output summary is empty. Thanks to this chart we could decide to set for all the experimentations CLD as the clustering method and α to 0.2.

Figure 5 shows the similarity values of the different methods when the default sizes of the summaries change. The minimum number of sentences in the content and in the context are both set to 1. Similarity values of the first method almost remain unchanged which may be explained by a straightforward reason: with the minimum degree of topicality set to 0.1, the average number of sentences in the summary when the default size is 5 is in reality nearly 3. Thus increasing the default size of the summary beyond 3 does not change the output summary. Clearly, context-based methods are more accurate than the content-based one provided that the sizes of the context and the content are not too small. The similarity values of the second method proves that the underlying hypothesis on which it is based on leads to more relevant summaries than the content-based approach.

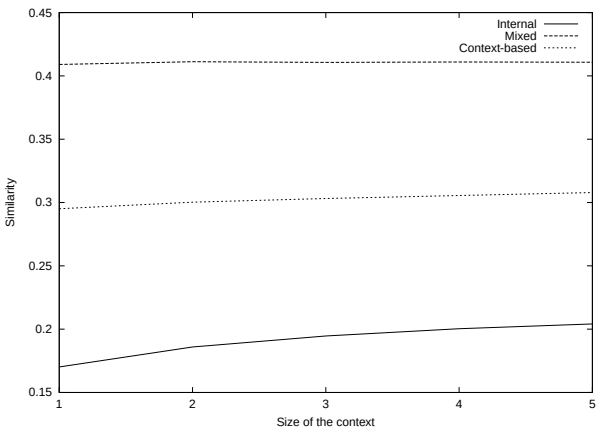

Figure 5: Sim. values when the summaries are resized

The following table sums up the results previously seen and the approach to choose with respect to the sizes of the context and the content of the target.

		Size of the context	
		≤ 4	> 4
Size of the content	≤ 3	internal	second method
	> 3	first method	first method

8. 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 new algorithms were proposed. Their efficiency depend on the size of the content and the context

of the target document. An evaluation technique derived from intrinsic evaluation techniques of summary was 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 on automatic tuning of the parameters of the two methods.

9. REFERENCES

[1] E. Amitay and C. Paris. Automatically Summarising Web Sites - Is There A Way Around It? In *Proceedings of the ACM 9th International Conference on Information and Knowledge Management CIKM*, pages 173–179, 2000.

[2] G. Attardi, M. S. Di, and D. Salvi. Categorisation by Context. *J.UCS: Journal of Universal Computer Science*, 4(9):719–736, 1998.

[3] A. L. Berger and V. O. Mittal. OCELOT: a system for summarizing Web pages. In *Research and Development in Information Retrieval*, pages 144–151, 2000.

[4] B. Bouchon-Meunier, M. Rifqi, and S. Bothorel. Towards general measures of comparison of objects. *Fuzzy Sets and Systems*, 84, 1996.

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

[6] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource list compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th International World Wide Web Conference*, 1998.

[7] B. D. Davison. Topical Locality in the Web. In *Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval (SIGIR 2000)*, pages 272–279, Athens, Greece, Jul 2001.

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

[9] J. Goldstein, M. Kantrowitz, V. O. Mittal, and J. G. Carbonell. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. In *Research and Development in Information Retrieval*, pages 121–128, 1999.

[10] S. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32:241–254, 1967.

[11] S. Jones, S. Lundy, and G. Paynter. Interactive Document Summarisation Using Automatically Extracted Keyphrases. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)*, volume 4, Big Island, Hawai, Oct 2002.

[12] F. Menczer. Links tell us about lexical and semantic Web content. Technical report, Computer Science, abstract CS.IR/0108004, arXiv.org, Aug 2001.

[13] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Five papers on Wordnet. Technical report, Cognitive Science Laboratory, Princeton University, 1990.

[14] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

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

[16] G. Salton, A. Yang, and C. Wong. A vector space model for automatic indexing. *Communication of the ACM*, 18:613–620, Jul 1975.

[17] H. Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[18] K. Sparck-Jones and J. Galliers. Evaluating Natural Language Processing Systems. *Springer*, 32:241–254, 1967.

[19] M. Zajicek, C. Powell, and C. Reeves. A Web navigation tool for the blind. In *Proceedings of the 3rd International ACM Conference on Assistive technologies*, California, 1998.

[20] Y. Zhang. World Wide Web Site Summarization, Master thesis. Technical report, Faculty of Computer Science, Dalhousie University, Apr 2002.