A multi-criteria Method for Automatic Web Page Summarization

Mehdi Belguith, Imen Touati, Mohamed Hédi Maâloul and Iskandar Keskes

ANLP-RG,MIRACL Lab.University of Sfax, Tunisia

Abstract . In this paper we propose an original method to automatically summarize Web pages. This method is based on statistics rather than linguistics. It differs from most other methods by its ability to generate summaries with textual content as well as non textual content (images/graphics). Our method consists of a multi-criteria analysis to determine the salient content to be considered in the summary. The method has been implemented and integrated into a fully automated Web page summarization system.

Keywords : Web pages, Automatic summarization, Multi-criteria analysis, Textual content, Non textual content, Selection criteria, Numerical approach.

1 Introduction

Summarizing a Web page is not an easy task. In the literature, several studies have considered the document summarization, but very few studies have addressed the Web page summarization. This is mainly due to the fact that web pages are not well structured as textual documents such as books, scientific papers, news articles, etc.

Initially Web page summarization methods were derived mainly from text summarization ones. However, it appeared that they were not effective to summarize Web pages. Indeed, there are several challenges to overcome: *i*) There is not a restricted domain on the Web and we can find everything from newspaper articles to lists of URL, *ii*) Punctuation marks are not often used in Web pages as in texts, *iii*) The Web pages may contain few words or portions of sentences that do not form a coherent text, and *iv*) The Web pages are multimedia, they may contain in addition to textual contents, non-textual contents (sounds, images, graphics, videos, links, etc.).

In this paper, we propose an original method to automatically summarize Web pages. This method is based on a multi-criteria analysis of textual and non-textual contents of Web pages. The criteria weighting is done in an objective manner using the Entropy method and the criteria aggregation is based on SAW (Simple Additive Weighting) method after normalization of the criteria scores. The entire proposed method has been implemented and evaluated through our Web-Summarizer system. A system demonstration is available on the Web[1].

[1] https://sites.google.com/site/websummarizer15/

This paper is organized as follows. In Section 2, we state the problem and present it in a formal way. Then, in Section 3, we propose 9 criteria to select the Web page relevant content to be considered in the summary. These criteria are classified into two classes: textual criteria (sentence position, title keywords, keyword frequency, sentence connectivity, "bonus" phrases, sentence length and formatting) and non-textual criteria (the criterion of a salient sentence pointing to an image/graphic and the criterion of an expressive image/graphic). We then detail in Section 4, the steps of the proposed method. In Section 5, we present the evaluation results of our WebSummarizer system. Finally, in Section 6 we give a conclusion and some perspectives.

2 State of the art

The state of the art distinguishes three main summarization approaches : the linguistic approach, the numerical approach that does not use any in-depth parsing and the hybrid approach that combines linguistic and numerical techniques (Nenkova and McKeown, 2011).

2.1 Linguistic approach

The linguistic approach produces summaries by comprehension (or by abstraction). It exploits techniques and models from artificial intelligence and cognitive psychology fields. Thus, the summary production has to go through a full or partial understanding phase. There are many methods using the linguistic (also called symbolic) approach based on understanding. (Blais, 2008) for example proposed a linguistic approach to discourse analysis of French texts in order to automatically generate a summary. To determine the discourse relations (Blais, 2008) used linguistic markers. In the same context (Keskes, 2015) proposed a method for automatic summarization of Arabic documents based on a deep analysis. This method consists of segmenting the text into discourse segments then it determines the semantic relations (i.e. discursive relations) between these segments according to the theory of the segmented discursive representation (SDRT). To generate the summary, the idea consists of selecting the segments that have relevant discourse relations and eliminate those with little relevance (e.g. presenting examples, hypothesis , etc.).

2.2 Numerical approach

The objective of this approach is to provide a summary rapidly, without any deep parsing. Indeed, this approach relies on surface (shallow) text analysis. Within this approach, there are two categories of methods: statistical and learning based methods.

Statistical methods.
Statistical methods generally involve computing scores for text segments (usually sentences). These scores are calculated based on several criteria ((Radev and Fan, 2000), (Bhatia et *al*., 2012), (Oufaida et *al*., 2014)). A sentence is then extracted if its

overall score is higher than a defined threshold. The main criteria taken into account in assessing the relevance of a sentence are the keywords frequency, the sentence position, the title keywords, some linguistic markers, etc.
What characterizes these statistical methods is that they use completely numerical values calculated using weights and scores.
(Liu et *al*., 2012) proposed a method for Chinese document automatic summarization based on compound words and keywords extraction. First, this method recognizes the compound words in a document and determines the part-of-speech to review the word segmentation. Then, it determines the keywords and calculates the sentences weights based on the keyword weights. Finally, it selects the sentences that have the highest weights in order to include them in the summary. According to (Liu et *al*., 2012) the generated summaries have good continuity and are understandable. The obtained average measures are 68.31 % for the precision and 66.72 % for the recall.
In the same context, (Boudin 2008) proposed a statistical approach for automatic summarization in the specialized field of organic chemistry. The summarization system of (Boudin 2008) has two modules. The first module applies a particular linguistic preprocessing sentences to take into account many specificities of organic chemistry documents. The second module selects the most important sentences based on a set of statistical criteria, some of which are specific to the chemistry field.
Other methods consist to analyze the contexts around the link of the Web page instead of the Web page itself, in order to generate the summary ((Kondratyev, 2005), (Chirita et *al*., 2006), (Jones and Building 2007), (Zhang et *al*. 2010), (Porselvi1 and Gunasundari, 2013)).

Machine learning based methods.
One could note the significant presence of machine learning approaches in the context of automatic summarization, based on texts/abstracts corpora. In the previous section, we saw that some criteria as the sentence position and the presence of certain keywords were used to determine pertinent sentences. Here, an important question arises: how to determine the contribution of each criterion in the sentence selection process ?
Of course, the answer to this question is dependent on the type of document to be summarized. Consider, for example, the sentence position criterion : in case of journalistic articles, first sentences are often the most important, while for scientific articles, sentences from the conclusion will be favored.
It is in this context that the learning approaches prove to be interesting. Indeed, the importance of each criterion can be estimated by counting their frequency in the corpus. Some researches has attempted to analyze how a corpus consisting of pairs (document/associated reference abstracts) could be used to automatically learn the rules for summary generation (Boudin 2008).
(Baratis et *al*., 2008) have proposed a machine learning method for image-based summarization of the Web. They choose the problem of summarization of large corporate Web sites by logo and trademark as a case study for the evaluation of the proposed method.
(Petinot et *al*., 2013) proposed a machine learning method to abstractive Web summarization based on the observation that summaries for similar URLs tend to be similar

in both content and structure. Two aspects of the graph have been trained, namely the edge templates and the slot locations.
(Bois et *al*., 2014) proposed the adaptation of the English language REZIME text summarizer to the French language. REZIME is a single-document summarizer particularly focused on summarization of medical documents. Summaries are created by extracting key sentences from the original document. The sentence selection employs machine learning techniques, using statistical, syntactic and lexical features which are computed based on specialized language resources.

2.3 Hybrid approach

The hybrid approach combines numerical and linguistic techniques. For example, (Zhang et *al*, 2010) proposed a method based on machine learning and automatic natural language processing techniques to automatically summarize Web sites. This method is based on four steps: extraction of URLs and texts, classification of narrative paragraphs, extraction of keyword phrases and extraction of relevant sentences. (Maaloul, 2012) has proposed an automatic summarization system for Arabic texts. This system is based on a hybrid approach which uses the RST (Rhetorical Structure Technique) to determine the rhetorical relations between the sentences. Then, sentences with important relations are selected for the summary. In case the system fails to detect the sentence relation, a learning technique is applied to determine whether the sentence is pertinent or not.

3 Multi-criteria analysis problem formalisation

We believe that the problem of choosing the salient sentences can be seen as a multi-criteria analysis problem.
Let P= $\{s_1, \ldots, s_n\}$ the set web page sentences. To choose the best sentences that will form the summary, we use a set C =$\{C_1, ..., C_q\}$ which constitutes a coherent criteria set. In order to judge the sentence pertinence according to each criterion, we define an evaluation function as follows : $C_j : P \rightarrow IR$

$s \rightarrow C_j(s)$

$C_i(s)$ represents the score of sentence s according to criterion C_i.
Thus, we calculate for each sentence s_i, a global score $GS(s_i)$ which represents the weighted sum of different scores of s_i according to all criteria:

$$GS(s_i) = \sum_{j=1}^{q} \alpha_j C_j(s) \qquad where \qquad \alpha_j > 0, \sum_{j=1}^{q} \alpha_j = 1$$

GS (s_i) is the global score of s_i and α_j is the weight of criterion C_j.

4 Proposed selection criteria

We present in this section the 9 selection criteria that we propose to select the relevant content of the Web page. We classify these criteria into two main classes, which correspond to the textual and non-textual web page content.

4.1 Textual content

For the selection of the textual content (relevant sentences) we use 7 criteria. These criteria are inspired from those used for texts but we propose to adapt them to the case of web pages, as we will deal with html files and not txt or doc file types.

Sentence position criterion (C_1).
Sentence position represents a criterion adopted by many research works dealing with automatic summarization. In these works, the sentences that are at the beginning are favored to those occurring at the end. Indeed, generally we pay more attention when writing the beginnings of texts, paragraphs, etc. We propose to calculate this score using the following formula: $\mathbf{C_1}(s)= m/pos(s)$

Where pos(s) represents the position of sentence s in the Web page and m, the number of sentences of the Web page.

Words titles criterion (C_2).
Titles or headings are important since they may contain relevant words. Thus, this criterion favors sentences containing words belonging to the titles and subtitles. The score assigned to the sentence according to this criterion is the number of title words that it contains: C_2 (s)= number of titles words of the sentence s.

Keywords criterion (C_3).
This criterion advantages sentences that contain keywords. To determine the web page keywords, we suggest to use the tf.idf technique (Term Frequency Inverse Document Frequency times) used in information retrieval to assign weights to the terms (words) of a document.
According to the tf.idf technique a word is important if it is relatively common in the web page and relatively rare in a large collection consisting of web pages linked by hypertext links to that web page.
We propose to ignore the "empty" words (e.g. conjunctions, pronouns, prepositions, etc.) that figure in a fixed list and calculate the tf.idf for the remaining terms using the following formula:

$$w_{ij} = tf_{ij} \times \log \frac{N}{n}$$

where w_{ij} is the weight of the term T_j in the page P_i ; tf_{ij} is the frequency of the term T_j in the page P_i ; N is the number of pages linked by hypertext links to the page P_i ; n is the number of pages where the term T_i occurs at least once.
We calculate for each word of the web page its tf.idf and we retain as keywords only those which tf.idf is above the average tf.idf.
The retained keywords are then enriched with the keywords that are in the keywords Meta tag (if it is available in the HTML file of the Web page).
The score of a sentence s, according to this criterion is the number of keywords contained in the processed sentence: $C_3(s)$ = number of keywords of the sentence s.

Sentence connectivity criterion (C_4).
In (Mani, 2001), the connectivity for a given sentence is defined by the number of sentences that are semantically related to it. To evaluate a sentence according to this criterion, we determine the number of sentences that contain words (other than empty

words) belonging to the considered sentence: $C_4(s)$ = number of sentences connected to the sentence s.

Bonus phrases criterion (C_5).

A "bonus phrase" represents a word or a group of words considered as important units as for example "the main objective ", "in conclusion ", "it's important to say ". Thus, this criterion advantages the sentences with one or many "bonus phrases":

$C_5(s)$= number of "bonus phrases" in the sentence s.

Note that we have defined empirically a list of 45 bonus phrases.

Sentence length criterion (C_6).

Generally, short sentences are preferred to long ones in the summaries. Thus, this criterion favors short sentences. We determine the average length (AL), in term of words, of the web page sentences using the following formula:

AL = sum of the sentences lengths / number of sentences.

AL is then used as a threshold to calculate the score of a sentence using the following formula: If Length(s) ≤ AL then $C_6(s) = 1$ otherwise $C_6(s) = 0$.

Formatting criterion (C_7).

According to this criterion, sentences with distinguished formatting such as a different color, size, style, or underlined, etc. are considered of a higher level of importance than normal ones. Thus, we propose three importance levels and we assign a different score to each level:

- Level 1: very important (score = 6)
- Level 2: important (score = 4)
- Level 3: somehow important (score = 2)

The determination of these levels is based on the two tag classes: Class 1 (<b>, <big>, , <font …>, <p…>, <div…>, <span…>) and Class 2 (<u>, <i>).

The importance level depends on the sentence tag class and is calculated as follows:

• If a sentence contains only tags of Class 2, it will have level 3 (i.e. $C_7(s) = 2$).

• If a sentence contains one or two tags of Class 1, it will have level 2 (i.e. $C_7(s) = 4$).

• If a sentence contains one or two tags of Class 1 and also tags of class 2 or a combination of three or more tags of Class 1, it is given level 1 (i.e. $C_7(s) = 6$).

Note that these values and these levels were chosen on the basis of an empirical study that we have conducted on a set of Web pages.

4.2 Non textual content

Given that an image/graphic can be expressive and can act as a summary in some cases, we can include them in the summary. For that we propose the "Image/graphic referring sentence" criterion and the "Expressive image/ graphic" criterion.

Image/graphic referring sentence criterion (C_8).

For this criterion, we propose to calculate the score of an image/graphic as follows:

A sentence that refers to an image/graphic (i.e. it contains a linguistic marker which refers to an image/graphic, as "the next image shows ...", "the following diagram indicates ...") and is followed by the image is advantaged to others and will have as a

score: $C_8(s) = 1$ otherwise $C_8(s) = 0$. Note that if this sentence is retained for the summary, it will be included with the correspondent image/graphic.

Expressive image/ graphic criterion (C_9).
This criterion concerns the images/graphics that are not referenced by any sentence. In this case, the image/graphic score is determined by the number of keywords contained in its description (i.e. the ALT attribute of the IMG tag).
In case the Alt attribute is empty, we use the hypertext link of the image/graphic to determine the number of keyword contained in this link.
Furthermore, if the image/graphic points to a web page, we consider the title of this page to determine the number of keywords (i.e. the title is used instead of the content of the ALT attribute when this latter is empty).
In all three cases, we obtain a total number of keywords MC(img) describing the image. The image/graphic score is given by the following formula:

$$C_9(\text{img}) = 1 \text{ if } MC(\text{img}) > 0 \text{ else } C_9(\text{img}) = 0$$

5 Criteria ponderation

It is obvious that the values of the criteria weights have a great influence on the ranking result in most aggregation methods.
There are several methods for determining the criteria weight. In this work, we opted for the Entropy method (Pomerol, 1992). This method is widely known in the literature of multi-criteria analysis since it is an objective method of criteria weighting that excludes any subjectivity of the decision maker in determining the criteria weights.
The idea is that a criterion j is particularly important when the variation between sentences scores, according to this criterion, is very important. Thus, the most important criteria are those that have the highest "power" of discrimination between sentences.
This method proposes to calculate, for each criterion, its entropy (Pomerol, 1992):

$$Ej = -K \sum_i C_{ij} \cdot \log(C_{ij})$$

Where C_{ij} is the score of sentence i according to criterion j.
K is a constant such that, for all j, we have $0 \leq Ej \leq 1$. For example, $K = 1 / \log(n)$ (n is the number of the web page sentences) is suitable.
The more the values C_j are close to each other the more the Entropy E_j is higher.
Thus, the weights will be calculated based on the dispersion measure (opposite of the Entropy): $D_j = 1 - E_j$. This weight will then be normalized by:

$$w_j = D_j \Big/ \sum_j D_j$$

6 Criteria Aggregation

In order to compare the sentences based on their respective scores, according to different criteria, it is necessary to apply an aggregation method of these scores. We propose to use the SAW method (Simple Additive Weighting) (Pomerol, 1992). SAW is an aggregation method which has the advantage to be simple and widespread. The principle of this method is to sum the obtained evaluations for each choice/ action (a

sentence in our case) according to the various criteria. For each sentence s_i, the global score (GS) is given by: $GS(s_i) = \sum_j w_j \cdot C_{ij}$

Where i=1,…, n ; j=1,…, q ; w_j is the weight of criterion j ; C_{ij} is the score of sentence s_i according to criterion C_j ; n is number of sentences ; q: number of criteria. We present in the following the different steps for criteria aggregation.

Step 1: Decision matrix construction
This step aims to construct the decision matrix : C = (C_{ij}) ; i = 1, ..., n; j = 1, ..., q that represents the respective scores of different sentences according to all criteria. C_{ij} is the score of sentence s_i according to criterion C_j .

Step 2: Decision matrix normalisation
This step consists to normalize the resulting decision matrix in order to make a homogeneous comparison across the different criteria that have different measure units. The elements of the decision matrix are normalized as follows:

$$Nij = \frac{C_{ij}}{\sqrt{\left\{\sum_{i=1}^{n} C_{ij}^2\right\}}},\ i=1,\ldots, n\ ;\ j=1,\ldots, q$$

Where : C_{ij} is the score before normalization ; N_{ij} is the score after normalization.
The obtained matrix after normalization is denoted by :

$$N= (N_{ij})\ ;\ i=1,\ldots, n\ ;\ j=1,\ldots, q$$

Step 3: Weighting of the normalized decision matrix
The normalized decision matrix is weighted and noted by:

$$V= (v_{ij})\ ;\ i=1,\ldots, n\ ;\ j=1,\ldots, q$$

It is obtained by multiplying each column of the normalized matrix by the relative weight of the criteria for that column. One element of this new matrix is determined by: $V_{ij}=\alpha_j N_{ij}$ where i=1,…, n ; j=1,…, q

Step 4: Sentence ranking
In order to rank the sentences, we determine the global score (GS) for each sentence according to the following formula:

$GS(s_i) = \sum v_{ij}$ where i=1,…, p and j=1,..,q

The sentence ranking is done according to the decreasing order of the sentence global score. Thus, the sentence with higher score is considered to be the most relevant.

7 Main steps of the proposed multi-criteria analysis method

Our method is based on four main steps to select the relevant content of a Web page.

Step 1: Web page pretreatment
The pretreatment step consists of cleaning the web page by removing the non-useful Meta tags, the script codes, the style sheets codes, the applets, etc. The aim is to retain only the tags that are useful for our method (i.e. that will be used in the calculation of the scores) such as the <title> and the formatting tags (Bold, font, I, strong, ...).

Step 2: Web page segmentation
The pretreated web page is segmented into headings, paragraphs and sentences based segmentation rules that rely on the punctuation marks and also on some HTML tags.

Step 3 : Sentence score calculation
After the segmentation step, we calculate, for each sentence, a global score which represents the sum of the normalized and weighted criteria scores. The computation of these scores follows the 4 steps described in section 6.

Step 4 : Summary content selection
This step consists to select important sentences and graphics/images that will appear in the summary. Thus, they are ranked according to the decreasing order of their global scores. Only sentences and images/graphics which scores are above a predefined threshold will be included in the summary.

8 Evaluation results

The entire proposed method has been implemented and evaluated through our Web-Summarizer system. A system presentation is available on the Web [1]. To evaluate it, we have compared 60 summaries generated by this system to the ones elaborated by a human expert. We have used a test corpus containing 60 Web pages from different domains as shown in table 3.2:

Theme	Web pages
Tunisian revolution	15
Club sportif sfaxien	10
Solar system	10
Tsunami earthquake	10
Computer Virus	15
Total	**60**

Table 3.2 Test Corpus

The obtained precision, recall and f-measure are respectively 66,3%, 64,5% and 65,38%.

9 Conclusion and perspectives

In this paper, we have proposed an original method for Web pages automatic summarization. This method consists of a multi-criteria analysis that considers both textual and non-textual contents of Web pages. We have defined 9 selection criteria where 2 of them corresponds to non textual content (i.e. images/graphics). The criteria are weighted in a purely objective manner according to the Entropy method. The entire proposed method has been implemented and evaluated through our Web-Summarizer system. The system demonstration is available on the Web[1]. The evaluation results are very encouraging. Indeed, the precision, recall and F-measure are respectively 66,3%, 64,5% and 65,38%.
As perspectives, we plan to extend this work by including, in the summary, an external image from the Internet in case there is no image in the Web page. This allows the

user to better understand the web page in a short time. We also intend to process web pages in other languages, such as Arabic and English (in addition to French). Note that our proposed method is language independent since it is based on statistical criteria. Indeed, we need only to build the specific lexicons for each considered language (empty words lexicon, etc.) and to identify the web pages segmentation rules.

References

(Baratis et al., 2008) Evdoxios Baratis, Euripides G.M. Petrakis, and Evangelos Milios, *Automatic Web Site Summarization by Image Content: A Case Study with Logo and Trademark Images*, IEEE Trans. Knowl. Data Eng. 20(9), 2008.

(Bhatia, et al., 2012) Kirti Bhatia, Dr. Rajendar Chhillar, *A Statistical Approach to perform Web Based Summarization*, IOSR Journal of Computer Engineering (IOSRJCE), Vol. 1, Issue 6, 2012.

(Blais, 2008) Antoine Blais, *Résumé automatique de textes scientifiques et construction de fiches de synthèse catégorisées : approche linguistique par annotations sémantiques et réalisation informatique*, Thèse de doctorat en Informatique linguistique, Soutenue à Paris 4, 2008.

(Bois et al., 2014) Remi Bois, Johannes Leveling, Lorraine Goeuriot, Gareth J. F. Jones, *Porting a Summarizer to the French Language*, 21ème Traitement Automatique des Langues Naturelles, Marseille, 2014.

(Boudin 2008) Florian Boudin, *Exploration d'approches statistiques pour le résumé automatique de texte*, Thèse en Informatique, Université d'Avignon et des Pays de Vaucluse, 2008.

(Chirita et al., 2006) Paul Alexandru Chirita, Claudiu S. Firan et Wolfgang Nejdl, *Summarizing Local Context to Personalize Global Web Search*, CIKM 2006.

(Jones and Building, 2007) Karen Sparck Jones and William Gates Building, *Automatic summarising: the state of the art*, Information Processing and Management, Special Issue on Automatic Summarising, 2007.

(Keskes, 2015) Iskandar Keskes, *Discourse Analysis of Arabic Documents and Application to Automatic Summarization*, Thèse de doctorat en informatique, Université de Sfax (Tunisie) et université Paul Sabatier (France), 11 mai 2105.

(Kondratyev, 2005) M. Kondratyev, *Web Sites Automatic Summarization*, Saint-Petersburg State University, 2005, http://syrcodis.citforum.ru/2005/kondratyev.pdf

(Liu et al., 2012) Xinglin Liu，Qilun Zheng, Qianli Ma, Guli Lin, *A Novel Automatic Summarization Method from Chinese Document*, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, N° 3, 2012.

(**Maaloul**, 2012) Mohamed Hedi Maaloul. *Approche hybride pour le résumé automatique de textes. Application à la langue arabe*. Thèse de doctorat en Informatique. Université de Provence - Aix-Marseille I, 2012.

*(***Mani, 2001)** Inderjeet Mani*, Automatic summarisation.* Amsterdam: John Benjamins,2001.

(Nenkova and McKeown, 2011) Ani Nenkova and Kathleen McKeown, *Automatic Summarization*, Foundations and Trends in Information Retrieval, Vol 5, No 2-3, pp. 103-233. 2011.

(Oufaida et al., 2014) Houda Oufaida, Omar Nouali, Philippe Blache, *Résumé Automatique Multilingue Expérimentations sur l'Anglais, l'Arabe et le Français*, 21ème Traitement Automatique des Langues Naturelles, Marseille, 2014.

(Petinot *et* al., 2013) Yves Petinot, Kathleen McKeown, and Kapil Thadani. *Cluster-based web summarization*. IJCNLP, Nagoya, Japan, 2013.

(Pomerol, 1992) J-C. Pomerol. *Choix multicritère dans l'entreprise : principes et pratique, HERMES*, 1992.

(Porselvi1 and Gunasundari, 2013) A. Porselvi1, S. Gunasundari, *Survey on Web page visual summarization*, International Journal of Emerging Technology and Advanced Engineering, Volume 3, Special Issue 1, January 2013.
(Radev and Fan, 2000) Dragomir R. Radev et Weiguo Fan, *Automatic summarization of search engine hit lists*, IRNLP 2000.
(Zhang et al, 2010) Yongzheng Zhang, E. Evangelos Milios, A. Nur Zincir-Heywood, *Topic-based web site summarization*. IJWIS 6(4), pp. 266-303, 2010.