AUTOMATIC WEBSITE SUMMARIZATION WITH IMAGE AND TEXT

[1]Prof. Vishal R Shinde, [2]Poojadevi B. Gupta, [3]Kiran A. Gaikwad, [4]Sagar R. Hankare
[1]Asst. Professor, [2,3,4]BE Student, [1,2,3,4]Comp. Engg. Dept, SSJCET, Asangaon, India.
[1]mailme.vishalshinde@gmail.com, [2]gupta.pooja.b@gmail.com, [3]kirugaik@gmail.com,
[4]hankaresagar73@gmail.com

Abstract: As the size and diversity of the World Wide Web (WWW) grows rapidly, Web sites become bigger and more complicated in content and structure and it is becoming more and more difficult to skim over their contents. This work is directed towards Web site summarization by image content focusing on the extraction of logo and trademarks from large corporate Web sites. This task is complementary to text summarization methods but, as opposed to methods that are based on text, the proposed method is based on image feature extraction from images and machine learning for distinguishing logo and trademarks from images of other categories (e.g., landscapes, faces). Because the same logo or trademark may appear many times in various forms within the same Web site, unique logo and trademark images are extracted first. These images are then ranked by importance. The most important Logos and Trademarks are finally selected to form the image summary of a Web site. The evaluation of the method demonstrated very promising performance.

Keywords — SAW (Simple Additive Weighting), RST (Rhetorical Structure Technique), segmented discursive representation (SDRT), arff (Attribute Relation File Format).

I. 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) Web pages are multimedia, they may contain in addition to textual contents, non-textual contents (sounds, images, graphics, videos, links, etc.). In this project, 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.

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.

II. LITERATURE SURVEY

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:

1) There is not a restricted domain on the Web and this project can find everything from newspaper articles to lists of URL.

2) Punctuation marks are not often used in Web pages as in texts,

3) The Web pages may contain few words or portions of sentences that do not form a coherent text.

4) This project, proposes 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.

The state of the art distinguishes three main summarization approaches:

1. The linguistic approach.
2. The numerical approach that does not use any in-depth parsing
3. The hybrid approach that combines linguistic and numerical techniques (Nenkova and McKeon, 2011)

LINGUSTIC METHOD

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. This method consists of segmenting the text into discourse segments[9] 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.).

NUMERICAL METHOD

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: 1.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,

HYBRID METHOD

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 . The problem of choosing the salient sentences can be seen as a multi criteria analysis problem. Let P= {s1, …,sn} the set web page sentences. To choose the best sentences that will form the summary, this project use a set C ={C1, ..., Cq} which constitutes a coherent criteria set. In order to judge the sentence pertinence according to each criterion, this project define an evaluation function as follows :Cj : P → IR s → Cj(s) Cj(s) represents the score of sentence s according to criterion Cj .

Thus, calculate for each sentence si , a global score GS(si) which represents the weighted sum of different scores of si according to all criteria.

	Summarization methods	concept	Advantages	Disadvantages	Application
1.	Linguistic	It produces summarization by comprehension. It exploits techniques and model form Artificial intelligence. It goes through a full or partial understanding phase.	This approach is done according to Thoery of segmented Discursive Representation (SDRT).	No other language can be handled.	Word-Net
2.	Numerical method	It produces summarization rapidly without any deep parsing. This approach relies on surface text analysis.	It is done using Rezime which is single document summarizer.	Cannot summarize multi documents of related topics.	Cue-word
3.	Hybrid method	It produces summarization by calculating sentences with important relations which are selected for the summary.	It uses RST (Rhetorical Structure Technique) to determine relations between sentences.	Data and words can be lost if they are repeated in paragraph or text.	Hurst parameter.

Table no.1.Comparison of summarization methods

III. MATHEMATICAL MODEL

The process of finding the most important Logos and Trademarks of a web site

IMAGE DETECTION

METHOD 1: CONTENT AREA INCREMENT

1st Circle Radius $r1 = \sqrt{\frac{\pi R^2}{\pi 256}} = \frac{R}{16}$

Where R the largest circle radius

Each circle radius: $r_1 = \frac{R}{16}\sqrt{\mathrm{i}}$

where i the enumerator of each circle (1-256).
dr definition: $\mathrm{dr} = \mathrm{r}_1 - \mathrm{r}_{\mathrm{i}-1}$.

METHOD 2: CONSTANT RADIUS INCREMENT

1st Circle Radius $r1 = \frac{R}{256}$

Where R the largest circle radius
Each circle radius: $r_1 = \frac{R}{256}\,\mathrm{i}$

where i the enumerator of each circle (1-256).
dr definition: $\mathrm{dr} = \mathrm{r}_1 - \mathrm{r}_{\mathrm{i}-1}$.

IMAGE COMPARISON:

$$m_{pq} = \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} i^p j\hat{}q$$

where i, j are the pixel co-ordinates. Translation invariance can be achieved by using the central moments:

$$\mu_{pq} = \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} (i-x)^p (j-x)\hat{}q$$

Where μ, are the co-ordinates of the regions center of gravity (centroid), which can be

$$x = \frac{m_{10}}{m_{00}},\ y = \frac{m_{01}}{m_{00}}$$

$$\theta_{pq} = \frac{\mu_{pq}}{(\mu_{00})^{\gamma}} \text{ where } \gamma = \frac{p+q}{2} + 1$$

rotation invariance is achieved with the seven invariant moments:

$$\phi_1 = \theta_{20} + \theta_{02}$$
$$\phi_2 = (\theta_{20} - \theta_{02})^2 + 4\theta_{11}^2$$
$$\phi_3 = (\theta_{30} - 3\theta_{12})^2 + (3\theta_{21} - \theta_{03})^2$$
$$\phi_4 = (\theta_{30} + \theta_{12})^2 + (\theta_{21} + \theta_{03})^2$$
$$\phi_5 = (\theta_{30} - 3\theta_{12})(\theta_{30} + \theta_{12})\left[(\theta_{30} + \theta_{12})^2 - 3(\theta_{21} + \theta_{03})^2\right] + (3\theta_{21} - \theta_{03})(\theta_{21} + \theta_{03})\left[3(\theta_{30} + \theta_{12})^2 - (\theta_{21} + \theta_{03})^2\right]$$
$$\phi_6 = (\theta_{20} - \theta_{02})\left[(\theta_{30} + \theta_{12})^2 - (\theta_{21} + \theta_{03})^2\right] + 4\theta_{11}(\theta_{30} + \theta_{12})(\theta_{21} + \theta_{03})$$
$$\phi_7 = (3\theta_{21} - \theta_{03})(\theta_{30} + \theta_{12})\left[(\theta_{30} + \theta_{12})^2 - 3(\theta_{21} + \theta_{03})^2\right] - (\theta_{30} - 3\theta_{12})(\theta_{12} + \theta_{03})\left[3(\theta_{30} + \theta_{12})^2 - (\theta_{21} + \theta_{03})^2\right]$$

The above seven invariant moments describe the images and are rotation-, translation-, and scale-invariant. Then Euclidian distance of invariant moments between pairs of images is computed. Similar images give small difference.

IV. PROPOSED SYSTEM

Proposed system works on text summarization, in this technique the unstructured text is converted into structured text. The second stage is to passed important key-phrases in the text by implementing the new algorithm through which extracting the high-frequency words.

The system uses the extracted keywords to select important sentences with highest rank from the input text.

The semantic similarity based on single document summarization. It takes two input parameters, the input text document and the no. of frequency terms. As the output generated a summarized text document along with the two measures compression ratio and the retention ratio. The single document text summarization

n is generated using frequent terms and semantic similarity.

Complete prototype system has been developed for the task of image-based web site summarization. System extracts its most characteristic images. These images form the image summary of the Web site. The purpose of this summary is then twofold:

(a) It is presented to the user for viewing and browsing.

(b) It can be stored and used by search engines for fast searching of the contents of the Web.

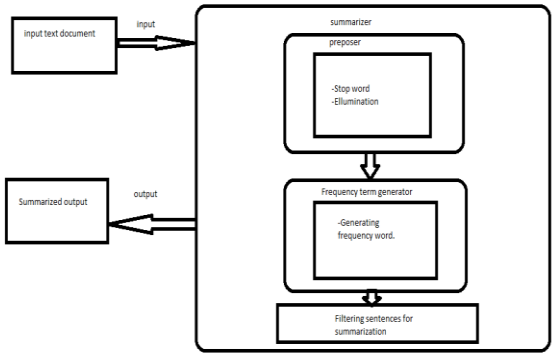

FIG 1: TEXT SUMMARIZATION

The method of the proposed system is implemented and integrated with a complete automated Website Summarization. The main focus is novel on the image feature extraction and important text sentence summarization

IMAGE FEATURE EXTRACTION

Logo and trademark are small size graphic images, with a limited number of intensity levels and colors. Image information is captured by Intensity and Histograms.

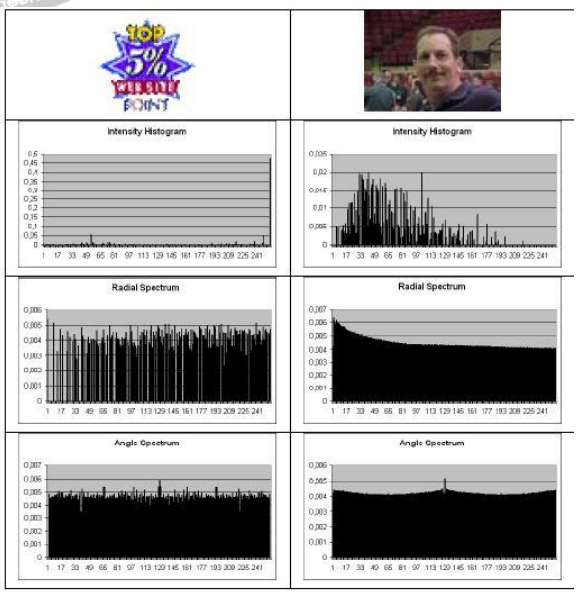

Fig 2: logo and non-logo images and of their Intensity and Frequency Histograms

LOGO AND TRADEMARK DETECTION

Machine Learning by Decision Trees is employed for Training the System to distinguish between logo and other images. This method estimates performance on data that has

not been used for training. It measures overall basis of probability of each image.

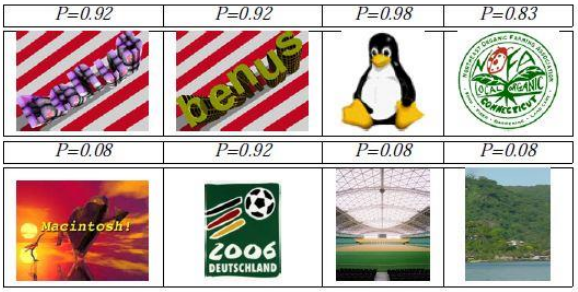

Fig 3: Images with their probabilities

IMAGE CLUSTERING

The purpose of this step is to group all the similar Images together into clusters. Because there may exist certain degree of uncertainty in detecting whether two images are similar.

Each cluster contains various instances of the same logo and image. From each cluster one image is selected to represent the cluster in the summery.

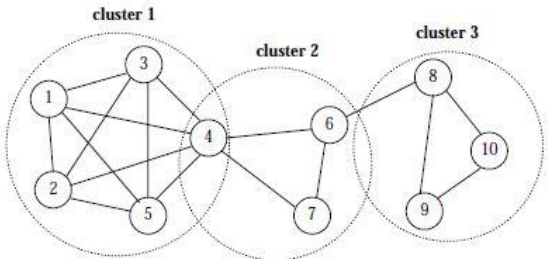

Fig 4: Clustering on the image similarity graph

IMAGE RANKING

The purpose of the step is to find most important images in the website. The importance of the image is computed based on the following criteria:

- Probability
- Instances
- Depth

Image Importance= Probability*Instances*Depth

Image based Summarization

The number of the clusters can be very large, and it becomes meaningful to be rank the clusters themselves by importance. The importance of the clusters depends on the importance of the Images it contains and is computed as

Cluster Image= $\sum_{i=image\ clusters}^{n}$ image Importance

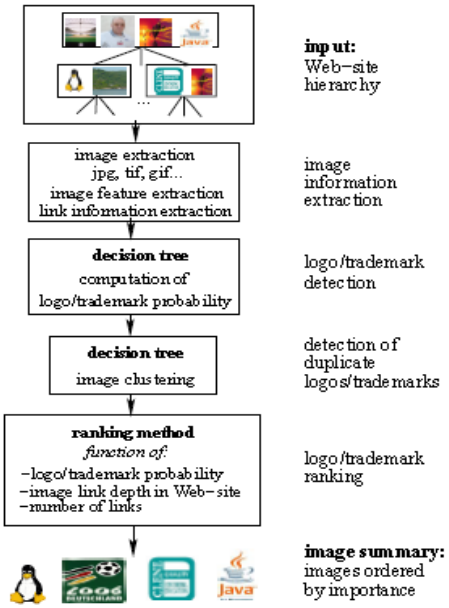

Fig 5: Image Ranking

A. AIM AND OBJECTIVE

The aim of the project is finding the most important Logos and Trademarks of a web site is divided into three steps.

- Training how to learn to extract logos and trademark images: This step is based on image feature extraction. These features describe the main characteristics of Logos and Trademarks.
- Clustering of similar images: There are cases where the same logo/trademark is used(displayed, pointed to by pages) more than once in a Web site. The same image may also appear with the different size, with the same or different color, or even as grey scale image.
- Image ranking: The final step includes the selection of the most important characteristic images from a Web site. This stage accepts the results of the previous stage and ranks images by importance.

The objective is Machine learning is used to discriminate between Logos and Trademarks and images of other categories such as person images, outdoor images, images of products etc.

Once all images have been extracted from a web page and the logo/trademarks have been detect, identical or similar images are grouped together into clusters. This step is also based on feature extraction and machine learning as the above.

B. SYSTEM ARCHITECTURE

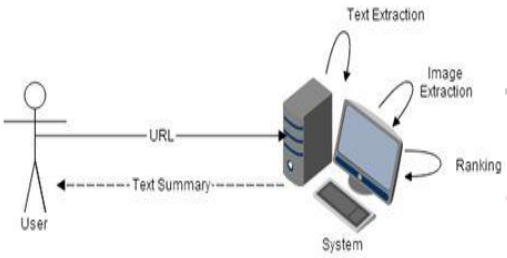

Fig 6: System Architecture

DESCRIPTION

User will give URL to system. System will extract the feature on web page. Ranking of sentences or image will be done. Depending on text or image, text summary will be generated. Ranking of system uses. Histogram for extracting image and for text used for the Frequent term list for extracting text from webpage. So this system will generate summary from image and text. If in webpage there is no text then it will extract images and generate summary.

V. ALGORITHMS

ALGORITHM FOR TEXT SUMMERY

Input: Text data from which summary is required.
Value of N for generating top N required terms
Output: Summary for original data Compression ratio.

Steps:
1.Data pre-processing phase
- Retrive data
- Eliminate stop word.

2 .For the entire text document
- Get the N frequent terms
- Generate term frequency-list.

3.For all N-frequent terms
- Generate sentences from the original data
- If the sentences consist of a term that is present in frequent-term-list then

4.Calculate compression ratio.

ALGORITHM FOR CLUSTERING THE SIMILAR IMAGES

Input: Features of pairs of images.
Output: Clusters of similar images.

1.For each pair of images
- .Compute their distance

2.Find pairs of images with distance less than T.
3.Compare with all other similar pairs.
- If they have at least one image in common merge and continue
- f it has been compared with all pairs, keep the cluster select the next pair and continue . At the end all the pairs with similar images are grouped together.

4.Output-The clusters of similar images.

ALGORITHM FOR GRAPH CLUSTERING

Input: Features for each image (mean value, threshold Standard deviation, energy, entropy).
Output: Decision trees

1.For each pair of images .
- Compute Euclidian distance of invariant moments, distance features and histogram intersections.
- .Assign the human defined class(similar or non-similar).

2.Form the above feature of arff file.
3.Pass the arff through the decision tree for training.

ALGORITHM FOR RANKING THE IMAGES.

Input: The cluster of similar images.
Output: The most characteristic-important images.

1.For each image in website.
- Compute the importance by

Importance=DEPTH*(BACKLINKS/ALL LINKS)*PROBABILITY
2.Sort images by importance.
3.Pick the most important image from each cluster.
4.Output the most important images of all clusters.

VI. CONCLUSIONS

A Web site summarization method focusing on image content is presented and discussed. We chose the problem of logo and trademark images as a case study for the evaluation of the proposed methodology. The problem of logo and trademark extraction (or Web site summarization by Logo and trademark extraction), is of significant commercial interest (e.g., Image Lock www.imagelock.com provides services on unauthorized uses of logos and trademarks) and this technology can benefit from the proposed approach. Extending the proposed methodology to handle any other image type is straightforward (i.e., the algorithms for logo and trademark selection, description and matching can be replaced by algorithms for the new image type).

The experimental results demonstrated that the method handles logo and trademark images and text summarization successfully in most cases and manages to extract the most characteristic images of this type from even from large corporate Web sites.

EVALUTION RESULTS

Web site	Number of images in Web site	number of logo & trademarks in Web site (human)	number of logo & trademakrs in Web site (method)	logo & trademark detection accuracy	logo & trademark classification accuracy	number of logo & trademarks in summary (human)	number of logo & trademarks in summary (method)	overall abstraction accruacy
www.berkeley.edu	100	2	1	50%	93%	2	1	50%
www.caltech.edu	35	3	1	33%	57%	3	1	33%
www.bu.edu	75	1	1	100%	92%	1	1	100%
www.cc.gatech.edu	25	3	3	100%	96%	3	3	100%
www.stanford.edu	122	7	6	86%	88%	7	4	57%
www.ocf.berkeley.edu	31	10	7	70%	87%	10	5	50%
www.umbc.edu	24	6	4	67%	92%	6	4	67%
www.debian.org	51	18	17	94%	98%	16	12	75%
www.java.com	19	4	3	75%	63%	3	2	67%
support.microsoft.com	84	10	8	80%	93%	6	5	83%
www.eclipse.org	54	5	4	80%	94%	6	3	50%
www.openoffice.org	38	6	4	67%	50%	5	4	80%
www.perl.org	26	25	20	80%	81%	22	14	64%
www.linux-france.org	15	3	1	33%	73%	3	0	0%
www.robocup2005.org	49	32	18	56%	71%	32	12	38%

REFERENCES

[1] Euripides G.M. Petrakis, E.V., EvangelosMilios, *Weighted Link Analysis for Logo and Trademark Image Retrieval on the Web*. .

[2] J.Sammon, M.S.L.O.G.M., *2. Global Image Analysis*, in *Practical Algorithms For Image Analysis. Description, Examples, and Code*, C.U. Press, Editor. 2000, The Press Syndicate Of The University Of Cambridge: Cambridge. p. 21-37.

[3] BorkoFurht, S.W.S., HongJiang Zhang, *4.3 Image Concepts And Structures*, in *Video And Image Processing In Multimidia Systems*, B. Furht, Editor. 1995, Kluwer Academic Publicers. p. 98-99.

[4] J.Sammon, M.S.L.O.G.M., *7. Frequency Domain Analysis*, in *Practical Algorithms For Image Analysis. Description, Examples, and Code*, C.U. Press, Editor. 2000, The Press Syndicate Of The University Of Cambridge: Cambridge. p. 246-264.

[5] J.Sammon, M.S.L.O.G.M., *3. Gray -Scale Image Analysis*, in *Practical Algorithms For Image Analysis. Description, Examples, and Code*, C.U. Press, Editor. 2000, The Press Syndicate Of The University Of Cambridge: Cambridge. p. 110-117.

[6] Flannery, W.H.P.W.T.V.S.A.T.B.P., *14. Statistical Description Of Data*, in *Numerical Recipes in C++ : The Art of Scientific Computing*. 2003. p. 616 - 617.

[7] Milan Sonka, V.H., Roger Boyle, *14. Texture*, in *Image Processing, Analysis, and Machine Vision*, K. McGee, Editor.1999, PWS. p. 646-653.

[8] Ian H. Witten, E.F., *3.2 Decision Trees*, in *Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations*, D.D. Cerra, Editor. 2000, Academic Press. p. 58-63.

[9] Ian H. Witten, E.F., *4.3 Devide and conquer: Constructing decision trees*, in *Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations*, D.D. Cerra, Editor. 2000, Academic Press. p. 89-97.

[10] Ian H. Witten, E.F., *6.1 Decision Tress*, in *Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations*, D.D. Cerra, Editor. 2000, Academic Press. p. 159-170.

[11] Ian H. Witten, E.F., *5.3 Cross-validation*, in *Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations*, D.D. Cerra, Editor. 2000, Academic Press. p. 125-127.

[12] Flannery, W.H.P.W.T.V.S.A.T.B.P., *2.6 Singular Value Decomposition*, in *Numerical Recipes in C++ : The Art of Scientific Computing*. 2003. p. 62-73.

[13] BorkoFurht, S.W.S., HongJiang Zhang, *11.2 Image Features For Content - Based Retrieval*, in *Video And Image Processing In Multimidia Systems*, B. Furht, Editor. 1995, Kluwer Academic Publicers. p. 230-232.

[14] Milan Sonka, V.H., Roger Boyle, *6.3 Region-based shape representation and description*, in *Image Processing, Analysis, and Machine Vision*, K. McGee, Editor. 1999, PWS. p. 259-262.

