ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

News Content Extraction from Web Content using PCA Classifier

Neha.M, Ancy Thomas

M. Tech, Department of CSE, Sapthagiri college of Engineering, Bangalore, India

Assistant Professor, Department of CSE, Sapthagiri college of Engineering, Bangalore, India

ABSTRACT: Web content extraction is a key technology for enabling an array of applications aimed at understanding the web. This project aims to extract less structured web content, like news articles, that appear only once in noisy WebPages. Our approach classifies text blocks by initially removing noise, then segmenting visual and text units by extracting features and PCA-based feature transformation for classification

KEYWORDS: WebPages, Visual Unit, Text Unit, Extracting Features, PCA.

I. Introduction

The World Wide Web (WWW) has been undergoing remarkable growth. Originated as a hypertext system for accessing many forms of documentation at CERN, the WWW rapidly grow as it is accessible for public use through the web browser. Along with its tremendous growth, the web has been experiencing many changes; one of them is related to how its content is presented to the user. Typically, a modern web document comprises of different kinds of content a news page, for instance, besides the article posting as the main content it also contains other noisy contents such as user comments, navigational menus, headers, footers, links to other news page, advertisements, copyright notices, privacy policies which scatter over the page. Considering the fact that a web document contains various forms of contents, it influences the way human browses the document. When browsing a particular web document, most of the time users typically focus on the main content and ignore the additional contents.

For human, this behavior can be done relatively fast and accurate because they can use their knowledge, visual representation and layout of the web pages to distinguish the main content from other parts. In the other hand, since computer software is not as intelligent as human to distinguish between the main content and the noisy content, this becomes the challenge for commercial search engines, web miners and other kinds of applications that use web document as a data source. A search engine, for instance, typically indexes the whole text of a web page. As a result, the noisy contents which is useless information remains in the index. The presence of noisy contents may degrade the performance of such Information Retrieval applications for example the quality of the search result, accuracy of information extraction, and the size of the index.

Many web applications can utilize the content structures of web pages. For example, some researchers have been trying to use database techniques and build wrappers for web documents. If a web page can be divided into semantic related parts, wrappers can be more easily matched and data can be more likely extracted. Link structure analysis can also make good use of the content structures of web pages. Links at different parts of a page usually act as different functions and contribute to the Page Rank or HITS differently. Recent works on topic distillation and focused crawling show the usefulness of page segmentation on information analysis. Furthermore, adaptive content delivery on small handheld devices also requires the detection of underlying content structure of a web page to facilitate the browsing of a large page by partitioning it into smaller units.

II. Related Work

Bar-Yossef et .al [1] define the common parts among web pages as template. When web pages are partitioned into some “pagelets” based on some rules, the problem of template detection is transformed to identify duplicated “pagelets” and count frequency. Their experiments show that template elimination improves the precision of the search engine Clever at all levels of recall.

Lin et.al [10] content-based approach is proposed by their system, InfoDiscover, partitions a web page into several content blocks according TABLE tags. Terms are extracted as features and entropy is calculated for each term and block entropy is calculated accordingly. An entropy-threshold is selected to decide whether a block is informative or redundant.

Yi et.al [11] make use of the common presentation style . A Style Tree is defined to represent both layout and content of a web page. Node importance is defined as the entropy of the node in the whole Style Tree for a site. By mapping a page of this site to the Site Style Tree, noisy information in the page is detected and cleaned. Their experimental results show that the noise elimination technique is able to improve data mining tasks such as clustering and classification significantly.

Gupta et al. [8] have proposed a DOM-based content extraction method to facilitate information access over constrained devices like PDAs. They implemented an advertisement remover by maintaining a list of advertiser hosts, and a link list remover based on the ratio of the number of links and non-linked words. But this rule-based method is relatively simple. For a portal web site like www.msn.com which is full of links, the rule would remove almost all useful content

III. PROPOSED SYSTEM

Figure 1 shows the architecture of proposed system. It includes following modules:

a. Noise removal:

In the first phase, web page pre-processing methods like html tag removal, tokenization, removal of stop words and stemming are applied on the input record and a feature set F of m tokens $\{x_{1,\ldots\ldots,}x_m\}$are retrieved. In the second phase, the HTML document is modeled as a DOM tree (Document Object Model tree). Each HTML page corresponds to a DOM tree where tags are internal nodes and the detailed texts, images or hyperlinks are the leaf nodes. We extract features like presentation style as well as the feature sets of individual blocks of the web page. In the third phase, a noise checking is done which is purely based on a feature set similarity measure. The weight percentage of each token in a feature set is calculated with respect to the total weight of the set and a new technique known as Minimum Weight Overlapping (MWO) is applied here for similarity verification. Calculated as the sum of minimum weights of each node. If it does not overcome a predefined threshold value t, that Leaf Node in the DOM tree is marked as a noisy node. This procedure is known as marking which is tried on all leaf nodes. By changing the value of t, we can control the relevancy of noise detection.

b. Visual Unit Detection:

The visual block extraction process is started to extract visual blocks of the current level from the DOM tree based on visual cues. Every DOM node is checked to judge whether it forms a single block or not. If not, its children will be processed in the same way. When all blocks of the current level are extracted, they are put into a pool. Visual separators among these blocks are identified and the weight of a separator is set based on properties of its neighboring blocks. After constructing the layout hierarchy of the current level, each newly produced visual blocks is checked to see whether or not it meets the granularity requirement. If no, this block will be further partitioned. After all blocks are processed, the final vision-based content structure for the web page is outputted.

c. Applying Segmentation Rules:

Based on our observation, there is a common pattern of text nodes that contain main content in a news or blog page. The text strings are always placed contiguously in the same depth level of a DOM sub tree. Most likely, the text strings are placed inside a p tag but however it also can be placed inside other tags such as div or plain #text node. The way we find the largest block of string is by traversing the DOM tree depth first based and then for each level in the sub-tree we find a child node with the longest string. After that we look to the sibling nodes of the child node in order to discover the neighboring text string. We will add the sibling's text string if it contains #text node. In order to skip the embedded noisy information such as advertisement or scripts, we check the content of the sibling's node. If it doesn't contain #text string as a child, we will skip it. We repeat this process until all the sub trees are examined. In the end, the output is the largest block of text string found in a sub tree. In case more than one DOM node is returned from the classification process, we may use this LBS heuristic to filter out noisy DOM nodes by comparing their longest string length.

d. Feature Extraction:

After the segment classification takes place, the output which are DOM nodes classified as good segment, are used as inputs for content classification. The content classification basically takes the feature vector of a segment and then classifies it whether it is a main content or noisy content.As we want to apply machine learning to do the segment classification and the content classification process, we need to transform the representation of the DOM node segments into feature vectors for classification tasks. The representation of our training data is represented as a feature vector:

$$< X_1, X_2, X_3, \dots, X_n, Y>$$

Xi The feature of the DOM node at indexes i in the vector.
Y The class label of the feature vector,

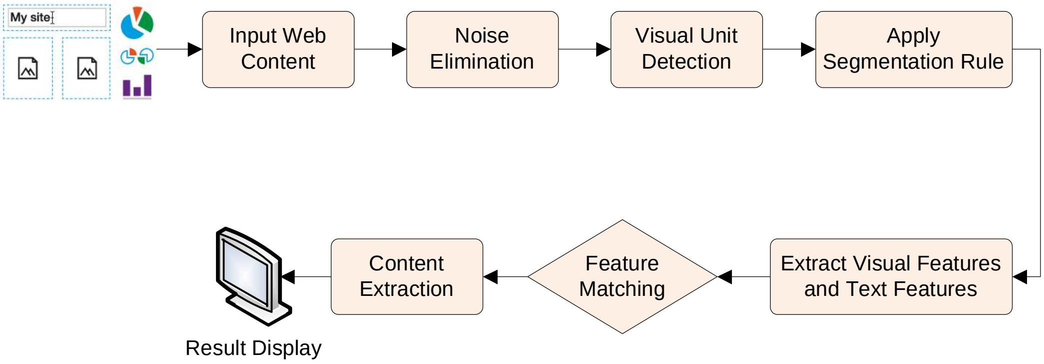


Figure 1: Proposed Architecture

a) Principal Component Analysis Based Classifier:

Principal Component Analysis Based Classifier Principal Component Analysis is a linear orthogonal transform method and is mostly used for dimension reduction in pattern recognition problems. It has also been used for dimension reduction in speaker recognition.

Principal Component Classifier can be briefly described as follows:

Let$X = [x_1, x_2, x_3 \dots, x_m]$denote the set of n dimensional feature vectors of k^{th} speaker. The sample covariance matrix of k^{th} speaker is calculated as

$$C = \frac{1}{M}(X - X^{-})(X - X) \qquad (1)$$

Where X is the sample mean of X. Let $(\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_n)$ be the eigenvalues ordered from the largest to smallest and $(w_1, w_2, w_3, \dots, w_n)$ be the associated eigenvectors of covariance matrix C. The principal component space is represented by the first r eigenvectors $p^{(k)} = [w_{k1}, w_{k2}, \dots, w_{kr}$ k=1,2,…,N, corresponding to the first largest r eigenvalues where N is the total number of speakers. The remaining(n-r)eigenvectors,$Q^{(k)} = [w_{k(r+1)}, w_{k(r+2)}, \dots, w_{kn)}]$ k=1,2,…,N, which are not chosen as principal components will be referred to as truncation error space. Generation of principal component space and truncation error space for each speaker is performed in the training step of speaker identification system

In the identification step, given a set of feature vectors, X, of an unknown speaker, the norm of the truncation errors of X, $\|Q^{(k)}(X - m_k)\|2$, k=1,2,…..,N is defined as the classification criteria. Finally, the unknown feature vector is assigned to the k^{th} speaker, of whom truncation error is the minimum, namely

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

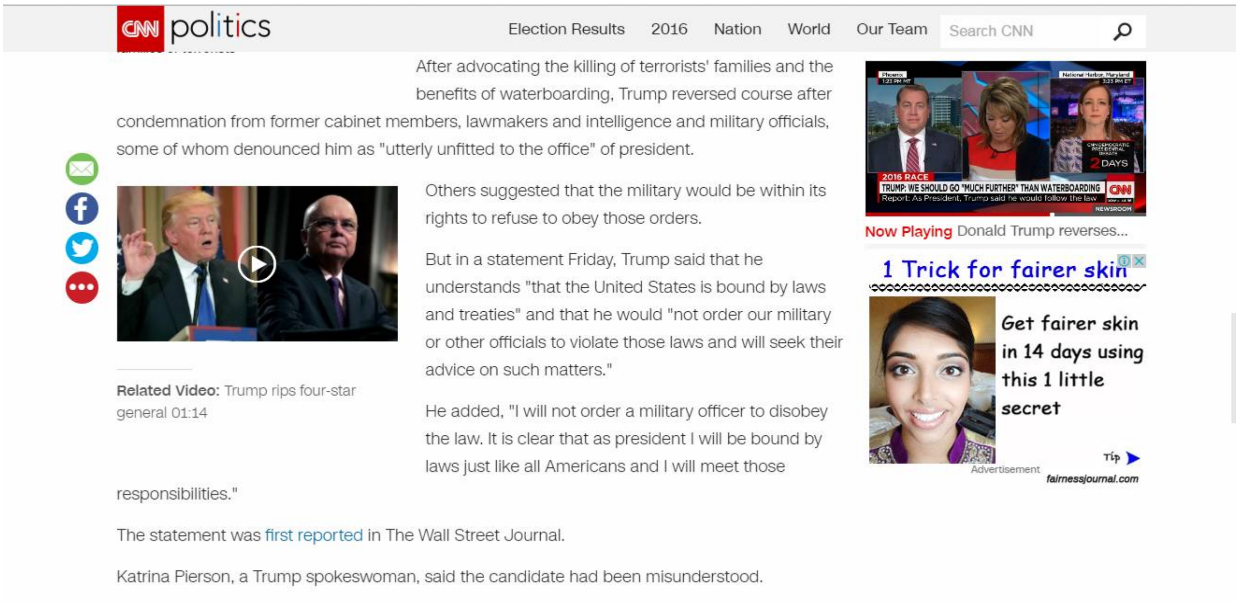
Vol. 4, Issue 4, April 2016

$$k^{\sim}=\arg\min\left\{\left\|Q^{(k)}(X-m_k\right\|^2\right\} \qquad (2)$$

b) EXPECTED RESULTS AND DISCUSSION

In this section explains the results of the proposed system.

Figure 2 shows input image to our proposed system. From this image we need to extract only the new contents excluding all other ads, images, and links. Figure 3 shows the result of segmentation performed for input image. In segmentation we have extract all the text regions from the input image for further processing. Final results are shown in Figure 4.


CNN politics
Election Results
2016
Nation
World
Our Team
Search CNN
2016 RACE
TRUMP: WE SHOULD GO "MUCH FURTHER" THAN WATERBOARDING
CNN
2 DAYS
Now Playing Donald Trump reverses...
1 Trick for fairer skin
Get fairer skin in 14 days using this 1 little secret
Advertisement
Tip
fairnessjournal.com

After advocating the killing of terrorists' families and the benefits of waterboarding, Trump reversed course after condemnation from former cabinet members, lawmakers and intelligence and military officials, some of whom denounced him as "utterly unfitted to the office" of president.

Others suggested that the military would be within its rights to refuse to obey those orders.

But in a statement Friday, Trump said that he understands "that the United States is bound by laws and treaties" and that he would "not order our military or other officials to violate those laws and will seek their advice on such matters."

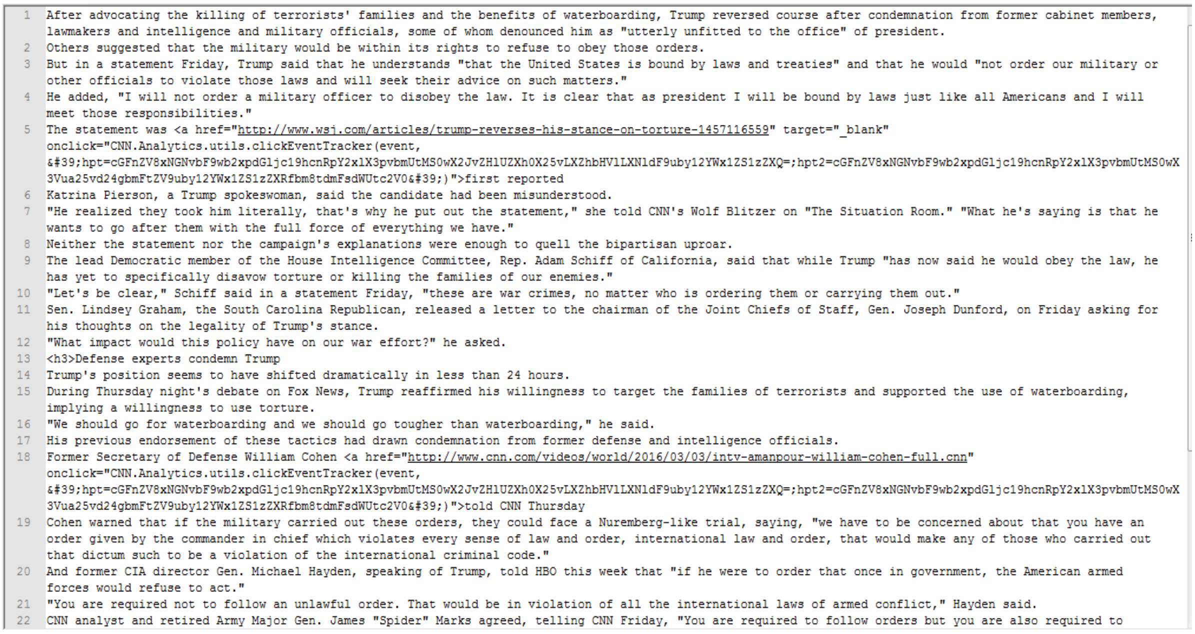
He added, "I will not order a military officer to disobey the law. It is clear that as president I will be bound by laws just like all Americans and I will meet those responsibilities."

Related Video: Trump rips four-star general 01:14

The statement was first reported in The Wall Street Journal.

Katrina Pierson, a Trump spokeswoman, said the candidate had been misunderstood.

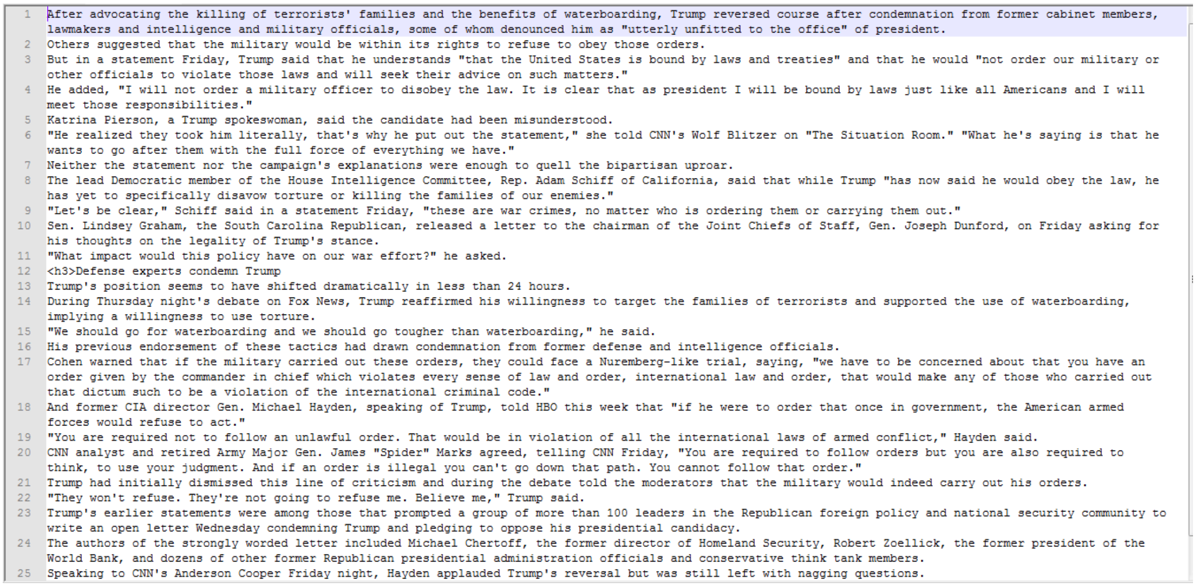
Figure 2: Shows the Input Image.



```
After advocating the killing of terrorists' families and the benefits of waterboarding, Trump reversed course after condemnation from former cabinet members,
lawmakers and intelligence and military officials, some of whom denounced him as "utterly unfitted to the office" of president.
Others suggested that the military would be within its rights to refuse to obey those orders.
But in a statement Friday, Trump said that he understands "that the United States is bound by laws and treaties" and that he would "not order our military or
other officials to violate those laws and will seek their advice on such matters."
He added, "I will not order a military officer to disobey the law. It is clear that as president I will be bound by laws just like all Americans and I will
meet those responsibilities."
The statement was <a href="http://www.wsj.com/articles/trump-reverses-his-stance-on-torture-1457116559" target="_blank"
onclick="CNN.Analytics.utils.clickEventTracker(event,
&#39;hpt=cGFnZV8xNGNvbF9wb2xpdGljc19hcnRpY2xlX3pvbmUtMS0wX2JvZHlUZXh0X25vLXZhbHVlLXNldF9uby12YWx1ZS1zZXQ=;hpt2=cGFnZV8xNGNvbF9wb2xpdGljc19hcnRpY2xlX3pvbmUtMS0wX
3Vua25vd24gbmFtZV9uby12YWx1ZS1zZXRfbm8tdmFsdWUtc2V0&#39;)">first reported
Katrina Pierson, a Trump spokeswoman, said the candidate had been misunderstood.
"He realized they took him literally, that's why he put out the statement," she told CNN's Wolf Blitzer on "The Situation Room." "What he's saying is that he
wants to go after them with the full force of everything we have."
Neither the statement nor the campaign's explanations were enough to quell the bipartisan uproar.
The lead Democratic member of the House Intelligence Committee, Rep. Adam Schiff of California, said that while Trump "has now said he would obey the law, he
has yet to specifically disavow torture or killing the families of our enemies."
"Let's be clear," Schiff said in a statement Friday, "these are war crimes, no matter who is ordering them or carrying them out."
Sen. Lindsey Graham, the South Carolina Republican, released a letter to the chairman of the Joint Chiefs of Staff, Gen. Joseph Dunford, on Friday asking for
his thoughts on the legality of Trump's stance.
"What impact would this policy have on our war effort?" he asked.
<h3>Defense experts condemn Trump
Trump's position seems to have shifted dramatically in less than 24 hours.
During Thursday night's debate on Fox News, Trump reaffirmed his willingness to target the families of terrorists and supported the use of waterboarding,
implying a willingness to use torture.
"We should go for waterboarding and we should go tougher than waterboarding," he said.
His previous endorsement of these tactics had drawn condemnation from former defense and intelligence officials.
Former Secretary of Defense William Cohen <a href="http://www.cnn.com/videos/world/2016/03/03/intv-amanpour-william-cohen-full.cnn"
onclick="CNN.Analytics.utils.clickEventTracker(event,
&#39;hpt=cGFnZV8xNGNvbF9wb2xpdGljc19hcnRpY2xlX3pvbmUtMS0wX2JvZHlUZXh0X25vLXZhbHVlLXNldF9uby12YWx1ZS1zZXQ=;hpt2=cGFnZV8xNGNvbF9wb2xpdGljc19hcnRpY2xlX3pvbmUtMS0wX
3Vua25vd24gbmFtZV9uby12YWx1ZS1zZXRfbm8tdmFsdWUtc2V0&#39;)">told CNN Thursday
Cohen warned that if the military carried out these orders, they could face a Nuremberg-like trial, saying, "we have to be concerned about that you have an
order given by the commander in chief which violates every sense of law and order, international law and order, that would make any of those who carried out
that dictum such to be a violation of the international criminal code."
And former CIA director Gen. Michael Hayden, speaking of Trump, told HBO this week that "if he were to order that once in government, the American armed
forces would refuse to act."
"You are required not to follow an unlawful order. That would be in violation of all the international laws of armed conflict," Hayden said.
CNN analyst and retired Army Major Gen. James "Spider" Marks agreed, telling CNN Friday, "You are required to follow orders but you are also required to
```

Figure 3: Shows the Segmented result for Input Image.

DOI: 10.15680/IJIRCCE.2016. 0404319



```
After advocating the killing of terrorists' families and the benefits of waterboarding, Trump reversed course after condemnation from former cabinet members, lawmakers and intelligence and military officials, some of whom denounced him as "utterly unfitted to the office" of president.
Others suggested that the military would be within its rights to refuse to obey those orders.
But in a statement Friday, Trump said that he understands "that the United States is bound by laws and treaties" and that he would "not order our military or other officials to violate those laws and will seek their advice on such matters."
He added, "I will not order a military officer to disobey the law. It is clear that as president I will be bound by laws just like all Americans and I will meet those responsibilities."
Katrina Pierson, a Trump spokeswoman, said the candidate had been misunderstood.
"He realized they took him literally, that's why he put out the statement," she told CNN's Wolf Blitzer on "The Situation Room." "What he's saying is that he wants to go after them with the full force of everything we have."
Neither the statement nor the campaign's explanations were enough to quell the bipartisan uproar.
The lead Democratic member of the House Intelligence Committee, Rep. Adam Schiff of California, said that while Trump "has now said he would obey the law, he has yet to specifically disavow torture or killing the families of our enemies."
"Let's be clear," Schiff said in a statement Friday, "these are war crimes, no matter who is ordering them or carrying them out."
Sen. Lindsey Graham, the South Carolina Republican, released a letter to the chairman of the Joint Chiefs of Staff, Gen. Joseph Dunford, on Friday asking for his thoughts on the legality of Trump's stance.
"What impact would this policy have on our war effort?" he asked.
<h3>Defense experts condemn Trump
Trump's position seems to have shifted dramatically in less than 24 hours.
During Thursday night's debate on Fox News, Trump reaffirmed his willingness to target the families of terrorists and supported the use of waterboarding, implying a willingness to use torture.
"We should go for waterboarding and we should go tougher than waterboarding," he said.
His previous endorsement of these tactics had drawn condemnation from former defense and intelligence officials.
Cohen warned that if the military carried out these orders, they could face a Nuremberg-like trial, saying, "we have to be concerned about that you have an order given by the commander in chief which violates every sense of law and order, international law and order, that would make any of those who carried out that dictum such to be a violation of the international criminal code."
And former CIA director Gen. Michael Hayden, speaking of Trump, told HBO this week that "if he were to order that once in government, the American armed forces would refuse to act."
"You are required not to follow an unlawful order. That would be in violation of all the international laws of armed conflict," Hayden said.
CNN analyst and retired Army Major Gen. James "Spider" Marks agreed, telling CNN Friday, "You are required to follow orders but you are also required to think, to use your judgment. And if an order is illegal you can't go down that path. You cannot follow that order."
Trump had initially dismissed this line of criticism and during the debate told the moderators that the military would indeed carry out his orders.
"They won't refuse. They're not going to refuse me. Believe me," Trump said.
Trump's earlier statements were among those that prompted a group of more than 100 leaders in the Republican foreign policy and national security community to write an open letter Wednesday condemning Trump and pledging to oppose his presidential candidacy.
The authors of the strongly worded letter included Michael Chertoff, the former director of Homeland Security, Robert Zoellick, the former president of the World Bank, and dozens of other former Republican presidential administration officials and conservative think tank members.
Speaking to CNN's Anderson Cooper Friday night, Hayden applauded Trump's reversal but was still left with nagging questions.
```

Figure 4: Shows the Final result for Input Image.

IV. CONCLUSION

Web news content extraction is vital to improve news indexing and searching in nowadays search engines, especially for the news searching service. In this paper we study the Web news content extraction problem and propose an automated extraction algorithm for it. Results show the accuracy of our proposed method.

REFERENCES

1. Nithya Bar-Yossef, Z. and Rajagopalan, S. " Template Detection via Data Mining and its Applications", in the proceedings of 11th World Wide Web conference (WWW 2002), May 2002.
2. Brin, S. and Page L. " The Anatomy of a Large-Scale Hypertextual Web Search Engine,"in the Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, 1998.
3. Cai, D., Yu, S., Wen, J.-R. and Ma, W.-Y., "VIPS: a vision basedPage segmentation algorithm," Microsoft Technical Report, MSR-TR-2003-79, 2003.
4. Chen, J., Zhou, B., Shi, J., Zhang, H.-J. and Qiu, F.,"Function-Based Object Model Towards Website Adaptation,"in the proceedings of the 10th World Wide Web conference Budapest, Hungary, May 2001.
5. Cover, T. M. "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," IEEE Transactions on Electronic Computers, Vol.14, pp. 326-334.
6. Dietterich, T. G. and Bakiri, G., "Solving multiclass learning problem via error correcting output codes," Journal of Artificial Intelligence Research, Vol. 2, pp. 263-286, 1995.
7. Dietterich, T.G. and Bakiri, G., "Error-correcting output codes: a general method for improving multiclass inductive learning programs," in the proceedings of AAAI-91, pp. 572-577.
8. Gupta, S., Kaiser, G., Neistadt, D. and Grimm, P "DOM based Content Extraction of HTML Documents," in the proceedings of the 12th World Wide Web conference (WWW 2003).
9. Budapest, Hungary, May 2003.Kovacevic, M., Diligent, M., Gori, M. and Milutinovic, V. "Recognition of Common Areas in a Web Page Using Visual Information: a possible application in a page classification. " in the proceedings of 2002 IEEE International Conference on Data Mining (ICDM'02), Maebashi City, Japan, December,2002.
10. Lin, S.-H. and Ho, J.-M., "Discovering Informative Content Blocks from Web Documents," in the proceedings of the ACM SIGKDD International Conference on Knowledge Discovery& Data Mining (SIGKDD'02), 2002.
11. Yi, L. and Liu, B., Web Page Cleaning for Web Mining through Feature Weighting, in the proceedings of Eighteenth International Joint Conference on Artificial Intelligence.(IJCAI-03), Acapulco, Mexico, August, 2003.
12. Yi, L. and Liu, B., Eliminating Noisy Information in Web Pages for Data Mining, in the proceedings of the ACM SIGKDD International Conference on Knowledge Discovery& Data Mining (KDD-2003), Washington, DC, USA, August,2003.
13. Yu, S., Cai, D., Wen, J.-R. and Ma, W.-Y., Improving Pseudo-Relevance Feedback in Web Information retrieval Using Web Page Segmentation, in the proceedings of Twelfth World Wide Web conference,2003.