LBNEST - Layout Based News Extraction and Syndication

S. Fernando, A.S. Perera, K. Wimalawarne, T. Buddhika and V. Kahawala,
Prabodha Fonseka, Thaminda Karunanayake

Abstract- **Layout Based News Extraction and Syndication is an attempt to implement a mechanism for extracting information from the news web sites. This research and development is carried out around some of the identified methodologies and finally the results obtained in each of the approach are discussed. Finally a hybrid approach is proposed which has been identified as the most accurate and efficient approach.**

I. Introduction

Layout Based News Extraction and Syndication project (a.k.a. LBNEST) is targeting on developing a flexible methodology for automating the news extraction from the news web sites. This project belongs to the area of layout based information extraction, which is subjected to a lot of research work in recent times.

With the internet boom, the need for information is increasing and internet has become the main source of information. Some modern concepts like *intelligent search engines* and *semantic web* emerged due to this information centered requirements. World Wide Web is considered as the dominating information source in the internet. But the web is still full of HTML based news pages containing a vast amount of information. For the concepts like intelligent search engines, there should be a way of extracting information contained in these web pages.

The requirement is addressed in LBNEST, but limiting the objective for extracting news from the news web sites.

This limitation is enforced, since it is impossible to develop a general mechanism for information extraction from the web due to its vast diversity of web sites.

As mentioned earlier, a lot of effort is put into research activities in the area of layout based information extraction. Few possible approaches have been identified for extracting information based on the layout of the web page. Proximity based approach; clustering based approach and DOM based approach are three of the possible approaches that can be used.

Proximity based approach focuses on the visual layout of the news webpage that needs to be processed; then extracts its news contents according to the visual layout of the news segments. The idea here is to identify the relationships between various segments (images, paragraphs, headlines, teasers etc.) in a news webpage by considering their visual propinquity.

Correlation clustering is a graph based information extraction technique. In this technique a graph, which having DOM (Document Object Model) nodes as the vertices, is constructed using the DOM tree of the HTML source.

DOM based approach basically deals with the HTML DOM of a web page and extracts information by processing the elements of the HTML DOM.

Each of these approaches possesses different advantages and disadvantages compared to each other. So depending solely on a single approach may not be the ideal solution. In LBNEST, the possibility of using these approaches is analyzed, and finally a hybrid approach is proposed which can be considered as the most effective approach.

II. Different Possible Approaches

A. Proximity Based Approach

This method focuses on the visual layout of the news webpage that needs to be processed; then extracts its news contents according to the visual layout of the news segments. The idea here is to identify the relationships between various segments (images, paragraphs, headlines, teasers etc.) in a news webpage by considering their visual propinquity.

Proximity Based Information Extraction can be used for two different purposes; one for direct information extraction and the other for improving the accuracy of an existing information extraction process. In this project we have used this method for the latter purpose due to several advantages that will be discussed later in the following sections.

There are five steps to follow in achieving the information extraction based on segments' proximity. First the news webpage should be extracted as HTML source. This is achieved by setting up a URL connection with the news webpage's URL and retrieving the input stream of the URL through that connection.

Even though we have retrieved the HTML source of the news webpage, there is no way to get an understanding about the webpage segments' proximities as the page is not rendered. Therefore, to determine the propinquities between the webpage segments, we should render the page using a rendering engine. Thus the second step is to render the extracted webpage using the parser provided by the 'Cobra HTML Library'.

After that an "RBlock", an object of the class that represents a tenderable block, can be attained from the parsed document. That is the block that represents the rendered webpage [4, 5]. We can retrieve the elements from the document as the third step.

As the fourth step in the process of Proximity Based Information Extraction we can separate the visual elements of the webpage by traversing through the node list retrieved. Traversal is done using the "Depth First Algorithm". The segments in the HTML source that actually possess visual areas when rendered can be cast in to the type "UINode". Then the size of the area occupied by the segment in the rendered webpage can be identified.

Even if the sizes of the rectangular areas are determined through this method, still the placement cannot be completed as the (x, y) co-ordinates of the elements are not absolute. If we plot the visual areas of the extracted webpage now, we'll get the boxes to overlap as illustrated in figure 01 (a) and (b).

Though the segment sizes are identified accurately through the getBounds() method, the (x, y) co-ordinates of the segments appear relative to their parent. Therefore, to place the segments in their original positions when we plot their respective rectangles, we have to add their ancestors' (x, y) co-ordinates to them.

Thus the fifth step is to calculate the absolute positions of the webpage segments. This can be achieved through two methods. One is to add the parents (x, y) co-ordinate values to their children when the children are identified. Other method is to add the ancestors' co-ordinate values to a node at the end by traversing back through the hierarchy.

The first method has proven to be effective than the latter. After the calculations are completed, the relationships between the segments can be determined by their proximities as closer segments are considered to be more related than the ones that are relatively distant. But none of these methods provides us with a 100% accurate layout of the webpage. The sizes of the webpage's visual areas are accurately provided but the (x, y) coordinates can be erroneous. There is no way to avoid these inaccuracies as the concrete classes provided by the cobra library produce them.

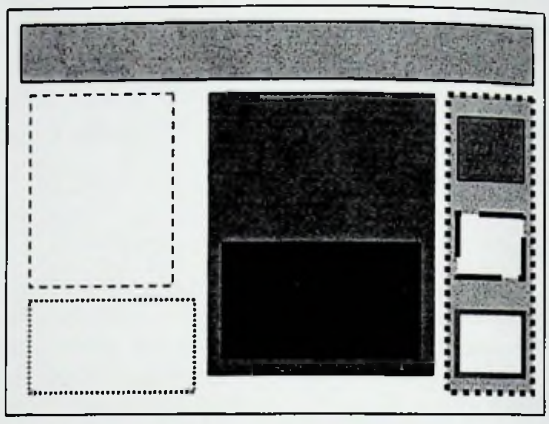

Fig. 1. Actual layout of the webpage

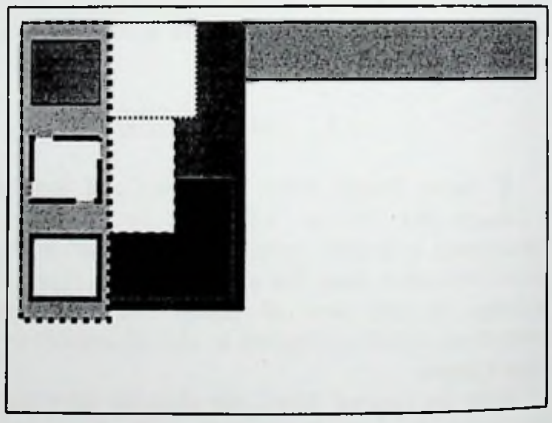

Fig. 2. Plotted layout of the webpage

There are advantages and disadvantages of this information extraction method. The main advantage of this method over the traditional methods like the DOM Based extraction technique is its high robustness [1]. The information extraction process of other methods is based on some assumptions on the HTML usage that include the names of the tags appearing in the neighborhood of the extracted content [3] or certain shapes of the document code tree [2]. However, considering the variability of the HTML language especially when used together with the Cascading Style Sheet technology, these assumptions generally cannot be ensured. This is the main cause of the low robustness of the traditional information extraction methods from HTML documents. As Proximity Based method renders the webpage on the run, no such issues arise.

The Proximity Based Information Extraction technique uses rendering of the extracted HTML source

hence increasing the running time. The traversal through the node list also affects the program's efficiency. That is one of the major disadvantages of this method compared to the other methods such as the HTML DOM Based method. Other than that, we can consider this method's inability to check the content of the segments as another disadvantage. This inability could lead into identifying advertisements displayed among the other news items as related news elements.

B. Clustering based Approach

Clustering based techniques can be applied to subdivide the web page in to segments and to group the related page segments in to a single cluster. One fine clustering technique which can be used in this case is "Correlation Clustering technique".

1) Correlation Clustering: This is a graph based segmentation technique. In this technique, a graph containing DOM nodes as the vertices is constructed using the DOM tree of the HTML source. The newly created graph is called the "Neighborhood Graph".

Neighbors of a particular node can be determined by the DOM tree distance or visual distance when rendering on the screen [6].

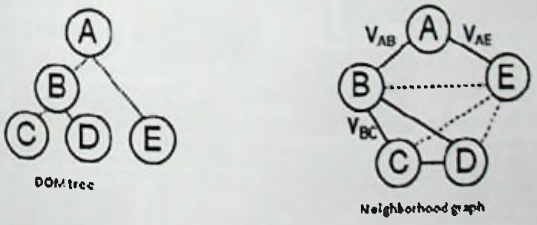

Fig. 3. Neighborhood graph

As illustrated in Fig.3 , the weight $V_{AB} \in [0, 1]$ of an edge represents the cost of placing its endpoints p and q in two different cluster; similarly, $(1 - V_{AB})$ represents the cost of placing p and q in the same cluster.

The weights between the nodes can be calculated by considering several properties of the DOM segment such as depth, height of the node, style rules of the node etc.

For example, if two particular nodes have the same height and an adjacent depth value, they have a high probability of being clustered in to the same cluster. This implies that most of the nodes having the same height have a higher chance to be in the same cluster, although the probability should be decreased while the depth increases. On the other hand, if two nodes have the same font color, line spacing, background color, font size, font weight etc., then those nodes have a higher probability of being in the same cluster and vice versa. Both dimensions and style rules of the node should be considered when calculating weights.

Following algorithm can be used to calculate the weights between the nodes. This is not an optimized algorithm and it was found in "trial and error method". By using this algorithm provided below, most of the unrelated images and site specific statements can be removed.

--Compare DOM node B with respect to node A. Assign a zero to initial variable say "initial-Cost".

--Take the height difference between two nodes. Assign 1.7 as the initial value for height (initial-Height). Iterate from zero to difference; deducting 0.07 from initial-Height in each iteration. At the end, add the final value to initial-Cost.

--If two widths are same in both documents, add 5.0 to the initial-Cost.

--Iterate through A's style rules and if a matching one is found in B's style rules, increment initial-Cost by 0.2.

--If the same style rule exists in both the nodes and if B is having a different value, deduct 9.0 from the initial cost.

--Divide final value by the X;

X = Math.sqrt (Math.sqrt (Math.sqrt (*A. height+ A. width*) * Math.sqrt (*B. height+ B. width*))))

--Final value is the similarity between the A node and the B node. The difference of two nodes can be calculated by deducting the final value from 1.

The algorithm can be further optimized by introducing the width of nodes. By this approach, the accuracy that can be obtained is about 65%. But with other approaches, a higher accuracy can be obtained, thus this approach is not used in the project.

Instead of calculating weights by an algorithm, there is another way to assign the weights to the neighborhood graph. A learning algorithm can be used to learn from a manually labeled data set. Properties of the data set also will be the dimensions, style rules and any other specific properties which can be used to identify a node among others. Since most of the training algorithms works with floating point numbers, all attributes should be converted to numerical format before used in a learning approach. In this approach classifying can be done with respect to the labeled data set having the labeled data set as a reference. Another approach is to just cluster the data set having the data set's attributes exactly as above. In this approach, attributes of the data set need not be numerical values. But there may be some error factor in this method, since it works without using a reference model. Calculated weights should be normalized into the range of zero to one in advance to perform clustering. The clustering algorithm is iterative. At each stage, a node p in the current graph is chosen uniformly at random and removed from the graph. A new cluster is created with just p in it. Next, all the nodes q such that $Vpq \geq \frac{1}{2}$ are removed from the graph, and placed in the cluster along with p [1]. The process is repeated on the remaining graph. Since the algorithm is randomized, several independent trials are performed and the solution with the least objective value is output as the answer.

2) Rendering Constraint: The rendering constraint arises out of the DOM tree structure: the area on the screen occupied by a child node is geometrically contained inside that of its parent node in the DOM tree. Therefore, any segmentation should respect the rendering constraint, i.e., if it places the root of a sub tree in a particular cluster, then it has to place all the nodes in the entire sub tree in the same cluster. [6]*C. DOM based Approach*In this approach, the HTML DOM (Document Object Model) is analyzed for the news extraction. During our research, at first we tried to solely use DOM based approach for news extraction and analyzed the accuracy of the results.

1) Jericho HTML Parser: For HTML DOM creation, an open source HTML library called Jericho was used. Jericho HTML Parser is a java library allowing analysis and manipulation of parts of an HTML document, including server-side tags, while reproducing verbatim any unrecognized or invalid HTML. It also provides high-level HTML form manipulation functions. It is an open source library released under both the Eclipse Public License (EPL) and GNU Lesser General Public License (LGPL). You are therefore free to use it in commercial applications subject to the terms detailed in either one of these license documents. [1]

Jericho was selected over the other open source HTML parsers, due to its provision of high level DOM manipulating functions which are similar to the XML parsers and in built functionalities to directly extract the standard HTML tags.

*2) Methodology used in the DOM based approach:*The methodology followed in the DOM based approach is depicted in the Figure 03.

Under this approach, both text based news items and related images are extracted from news web sites. Each activity in the above sequence of activities is discussed below with the code samples whenever necessary.

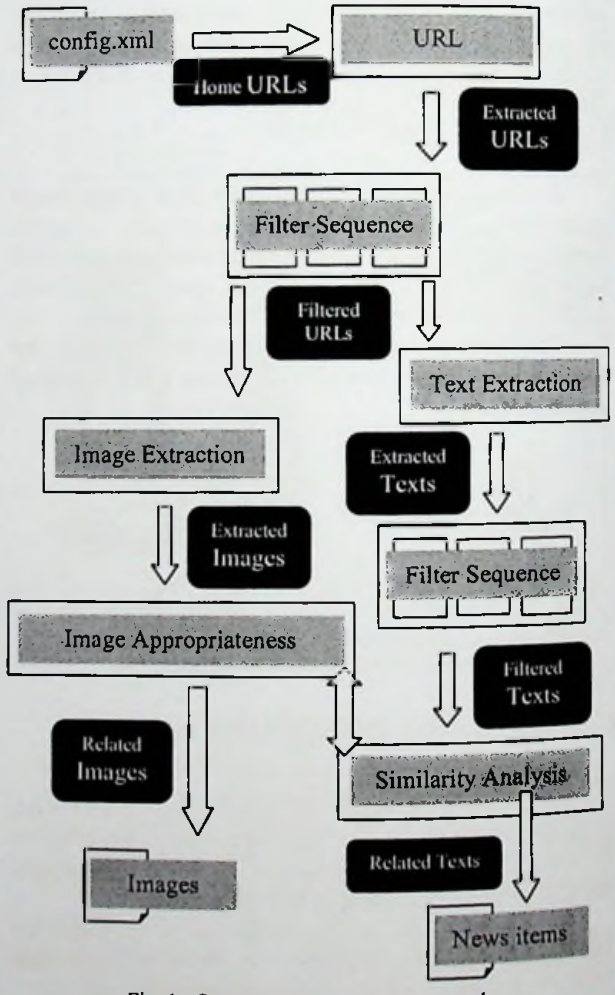

Fig. 4. Steps of the DOM based approach

i.URL Extraction

Most of the news web sites are using their home pages for displaying the headlines of the news items and corresponding links to the full stories. They are following this strategy to increase the news coverage of the home page. In this phase of the application, the URLs contained in the home pages of the news web sites are extracted. Home page URLs for the news web sites

are specified in the XML based configuration file. All the extracted URLs along with their anchor texts (if available) are pumped to the URL filtering phase.

ii.URL Filtering

All the hyperlinks displayed on a home page of a news web site are not a set of pointers to news stories, rather a collection of URLs for news stories, advertisements and certain site specific information. So it is necessary to filter these URLs and separate the URLs that are pointing to news stories. For this purpose, filtering is used in LBNEST.

In LBNEST, it is possible to declare filter sequences and use them for filtering strings. (Filtering is discussed in detail later in this section) There is a set of custom filters implemented for filtering URLs to remove the advertisements and site specific hyperlinks. These filters are implemented based on some heuristics. Word count of the anchor text and the similarity between the anchor text and the site name are two heuristics used. If the word count of the anchor text is less, then the URL rarely points to a news story. The threshold word count used in the LBNEST is four, for rejecting an URL. Also the high similarity between the news site name and the anchor text also indicates that the link is pointing to a site specific page. Similarity Analysis module is used for estimating the similarity between the anchor text and the site name. (Similarity Analysis module is discussed in the following sections)

If an URL does not contain a sufficient word count, it is further analyzed to check whether it is pointing to a news web site. Some news web sites represent news items as a combination of a teaser and a link to the full story with an anchor text like "Read more" or "Full Story". This is common in most of the local news web sites. Global news web sites like BBC and CNN do not follow this pattern when publishing news.

Indo-Lanka relations never so warm and close - Menon to President

[January 17 2009]

The relations between India and Sri Lanka have never been so close, so warm and so deep, said Indian Foreign Secretary Shiv Shankar Menon at his meeting with President Mahinda Rajapaksa at Kandy this morning (17).

» Full Story

Fig.5. An example of a teaser and a full story link

In such situations filtering based on the anchor text does not work. So when an URL is rejected from the word count based filter it is forwarded to a separate routine that seeks for the closest text element to the URL in the HTML DOM. This routine is implemented by considering the TREE aspect of the HTML DOM. The closest text element for a certain URL element is searched in the order of the levels of neighbor elements defined for that URL element.

Fig. 6 depicts the different levels of neighbor elements associated with a particular URL element. These levels are defined based on the number of steps required for reaching that element from the element of interest.

But it is should be verified that the text found in this approach is the corresponding teaser. Once the closest text is identified the corresponding page pointed by the URL is fetched and similarity analysis is done with the content of that page against the identified text. If a reasonable similarity is resulted, then URL is not rejected.

iii.Text Extraction

For each URL which gets qualified from the above phase, a text extraction is done. The content of the "p" tags are extracted in this phase.

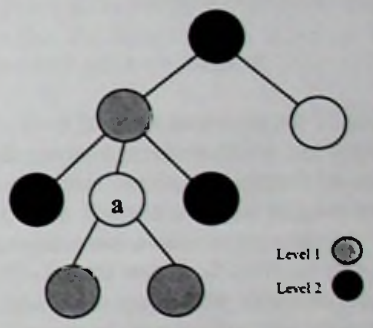

Fig. 6 Levels of neighbor elements

iv.Similarity Analysis

Since the news web pages contain other information, other than the news stories, the extracted text should be purified. For this purpose, similarity analysis is carried out to extract only the texts that are relevant to the news story. The anchor text or the teaser is compared with the extracted text and a similarity ranking is calculated. This similarity analysis is done in a sentence vice manner. (More about the similarity analysis module can be found later in this section.)

v. Image Extraction

Image extraction takes place just after to the text extraction process. The web page pointed by a filtered URL is sequentially searched for an image tag. When an "img" tag is encountered the closest text element is searched in a similar way as discussed in the URL extraction section but with a slight adjustment. In this case the HTML DOM is traversed only to the bottom,

where as in the previous case the DOM was traversed in both directions. This method is followed in order to avoid the effect of the images appearing in the footer of the web page. Normally news web pages contain thumbnails of images that correspond to the other news stories along with their headlines at the footer. If the previous approach is used for those images then there is a possibility of getting a text that is part of the news story as the closest text due to the upward traversal. Once a text element is found, it is checked with the extracted news story to check whether it is part of the news story. For this step to be successfully executed, the text extraction and similarity analysis on extracted text should be completed in advance.

The above mentioned approach is capable of precisely filtering out only the relevant image and leaving out the irrelevant images.

At the end of this sequence, a set of news stories are resulted with the relevant images.

III. Similarity Analysis

One of the important areas of news extraction and filtering is text based similarity analysis. Especially for the task of testing relatedness between different segments spread over different places in news pages, but belongs to the same news article, we mainly used text based similarity analysis. So in next section of this paper we focus on available techniques for analyzing text similarity and the approach we used for our task.

A. Similarity Ranking Algorithm

This algorithm, introduced by Simon White, is governed by the following requirements [8].

--A true reflection of lexical similarity – Strings with small differences should be recognized as similar. A significant substring overlap should point to high rank of similarity

--A robustness to change of word order – Two strings which have different word sequences should be recognized as similar

--Language Independence – The algorithm should be used not only for English language but for many other languages.

The similarity metric of this algorithm rewards both common substrings and common ordering of those substrings. It also considers not only the longest common sub string but also other common sub strings too. This purpose is achieved by finding out how many adjacent and equal character pairs are contained in both strings. Since each character pair contains a little information of the original ordering of the characters in the string, in addition to the number of adjacent characters, the character ordering of the original string is also considered. The algorithm can be described using the words 'France' and 'French'.

First both the strings are converted to their upper case letters to make the algorithm case insensitive. Then the two strings are split into their character pairs [8].

S1 = FRANCE: {FR, RA, AN, NC, CE}
S2 = FRENCH: {FR, RE, EN, NC, CH}

Then the similarity ranking between two strings S1 and S2 can be computed as follows

$$\text{Similarity}(S1, S2) = \frac{2 \times (\text{number of common character pairs in S1 and S2})}{\text{Total number of character pairs in strings S1 and S2}}$$

For the above example this would be

$$\frac{2 \times |\{FR, NC\}|}{|\{FR, RA, AN, NC, CE\}| + |\{FR, RE, EN, NC, CH\}|}$$

$$= (2 \times 2) / (5+5) = 0.4$$

B. Matching a News Headline with its Teaser or the Story

In our approach, news story or teaser was extracted as separate paragraphs. We checked the relatedness of each paragraph of each paragraph with the relevant news headline by breaking the paragraph into separate sentences and checking their similarity ranking values with the news headline. Analyzing the test results we obtained by comparing many sentences, we concluded that a similarity value greater than 20% is sufficient for two sentences to be related.

Even if a sentence does not have a significant level of similarity, it can be considered as a related sentence if its adjacent sentences have a higher similarity. Before performing the similarity analysis, all the texts were filtered to remove the common words in English. In addition to that, word stemming was also applied on the extracted texts.

IV. HYBRID APPROACH

In our process of news extraction we have employed a hybrid approach of the above mentioned DOM based and Proximity based approaches. As mentioned before, we have applied the Proximity based information

extraction approach to improve the accuracy of the DOM based approach.

A news webpage that displays a full story contains a lot of website specific information and other undesired links which have no relevance to the story presented in the page. Therefore, it will leave the DOM based method to do a lot of work to retrieve the desired news item (the full story) from the whole webpage. In a news webpage that contains a full story, the majority of the page's visual area is utilized by the text that includes that story. Therefore, when our requirement is to extract just that story text from the webpage, it will be convenient if we could apply the DOM based method just to the area that is of our concern.

This is where we have brought the Proximity based approach in to play. Through that approach, we identify the visual area of the webpage that contains the full story and then apply the DOM based technique to extract information from that area. That prevents the DOM based technique from processing unnecessary information (advertisements, website specific information, links to other news items etc) displayed in the webpage thus improving the accuracy of the overall extraction process.

The basic assumption made in using the Proximity based technique to separate the full story part from the webpage is, that the full story text possess the largest single visual block of that particular webpage. Therefore, our target is to identify the largest visual block in the webpage. We consider only the visual areas that contain text inside that block, eliminating the possibility of choosing an image as the story required.

The algorithm also verifies whether the visual block chosen as the one containing the full story has its (x, y) co-ordinates - co-ordinates of the block's top left corner - are significantly distant from the top left corner of the webpage - where the co-ordinates are (0, 0). This check is carried out with the assumption that a news website would never start its full story from the top left corner of the webpage.

We have tested this algorithm for different full story news web pages from several news websites and every time the results were accurate. The algorithm will first find the visual block that contain the full story and will send its contents to the DOM based technique to extract the full story from that block. If the conditions are not met (i. e. if the largest block has its (x, y) co-ordinates as (0, 0)), the whole webpage will be sent to the DOM based technique to process it and find the full story. Therefore, this hybrid implementation will never result in a data loss when the conditions are not met, but will certainly increase the overall process's efficiency when the conditions are met.

V. Identifying The Similar News Components

The approach is to identify similar news items, which are extracted from various news websites, and store them in separate clusters considering their similarity. These types of clustering techniques can be considered as natural language clustering. There are several techniques that can be used to perform this type of clustering. Through this document, how to use hierarchical clustering in order to perform natural language clustering is explained.

A. Hierarchical clustering

In hierarchical clustering, approach is to start from assigning each data item in to a single cluster. For example if there were n number of data items, there should be n number of clusters initially. After the initial assignment, algorithm is performed iteratively, in order to minimize the number of clusters. There are several ways that the algorithm can minimize the number of clusters.

1) Single link clustering :Similar pairs of data items are selected in such a way that, the distance between one cluster and another cluster to be equal to the shortest distance from any member of first cluster to any member of the second cluster [10].

2) Complete link clustering:Similar pairs of data items are selected in such a way that, the distance between one cluster and another cluster to be equal to the longest distance from any member of first cluster to any member of the second cluster [10].

B. Clustering analyzing terms [11]

In order analyze a clustering algorithm, there should be set of reference clusters (reference partition), to which the newly created clustered (response partition) can be compared. There are few measures that should be aware before analyzing the clustering algorithm.

--True positives – when pair of data item is taken, if those two exist in the single cluster in the reference partition and in single cluster in response partition, then that pair is a true positive.
--True negatives – if the two data items don't exist in a single cluster in either of the reference or response partitions.
--False positives – if the two data items exist in two different clusters in reference partition and single cluster in the response partition.
--False negatives – if two data items exit in a single cluster in reference partition and two different clusters in response partition.

Example:

Reference partition: {[1, 2, 3], [4, 5], [6]}

Response partition: {[1, 2, 3, 4], [5, 6]}

TABLE I
RESULTS TABLE FOR ABOVE EXAMPLE

True positives	True negatives	False positives	False negatives
(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)	(1,4), (4, 1)	(3, 4), (4, 3)	(1, 3), (3, 1)
(1, 2), (2, 1)	(1, 5), (5, 1)	(3, 5), (5, 3)	(2, 3), (3, 2)
(4, 5), (5, 4)	(1, 6), (6, 1)	(3, 6), (6, 3)	
	(2, 4), (4, 2)	(4, 6), (6, 4)	
	(2, 5), (5, 2)	(5, 6), (6, 5)	
	(2, 6), (6, 2)		
10	12	10	4

Following measures can be used to evaluate the clustering algorithm [12]

Accuracy = Correct responses / Total = (True positives + True negatives)/ Total

Precision= True positives / (True positives + False positives)

Recall= True positives/ (True positives + False negatives)

C. Distance calculation

Whatever the clustering technique is, the most important thing in any kind of clustering is distance calculation. The following distance calculation algorithm works with the word occurrences in a document. Algorithm is as follows [11].

--Break the document into tokens which are separated by a space.
--Remove full stops, commas, and common verbs such as "is, was, were, etc" from the token list.
--Get the number of tokens in the document and get the square root of the number. This is the length of the document.
--Get a count which representing the number of occurrences of a particular word. For example if "town" word is repeated in the document four times, the town count is four for that document.
--When comparing document A with document B, iterate through A's tokens and if B contains a token, get the two counts, which represent occurrences of the token in two documents, multiply them and get the square root.
--Repeat above instruction until A's tokens are all over and get the sum of every square root values.
--Take the ration of the value returned from the above instruction over multiplication of two document's lengths. This is the similarity of the two documents with a fine accuracy. One minus the similarity value is the distance between two documents and this can be used to cluster documents.

A good threshold value can be obtained by analyzing several data sets. Measures such as Accuracy, Recall, and Precision can be used to determine the value. This analysis can be in two different ways. One is to identify number of clusters which gives the most accuracy with respect to a reference partition. Other one is to identify the threshold distance value, which can be used to cluster using hierarchical clustering. In this approach data items, which are having lesser distance that the threshold, are grouped in to a single cluster depending on their inter dependencies. For example if the cutoff distance is given as 2, and if A, B are having distance of one in between them and C and D are having distance of one in between them, A, B, C and D are clustered together, if the distance between A, C or A, D or B, C or B, D are lesser than 2. If none of above is true A, B and C, D are clustered as two clusters.

In the project, in order to cluster news stories, a threshold value is used. Threshold values between two similar news items are identified as 0.35 to 0.4. For the first time 85 news items were clustered and 69 clustered were obtained with the 0.38 threshold value. All the clustered news stories had a relationship when considering the meaning of them.

VI. ANALYSIS OF RESULTS

After the completion of the implementation phase, some test runs were carried out for analysis purposes. Similarity Analysis module and URL filtering mechanism were tested as individual components whereas the overall system was tested for both the DOM based and hybrid approaches.

A. Results for URL filtering mechanism

For this test, three popular news sites were chosen, namely BBC, CNN and Sky News. Then the samples of 80 URLs were extracted from each site containing 30 site specific URLs and advertisement URLS with the rest pointing to news stories and fed into the filtering sequence. Filtering was done based on either the anchor

texts or teasers. Collected statistics are tabulated as below.

TABLE II
RESULTS OF URL FILTERING

Site	URLs for news stories		URLs for non news items		Accuracy (%)
	Accepted (/50)	Rejected (/50)	Accepted (/30)	Rejected (/30)	
BBC	47	3	2	28	93
CNN	43	7	3	27	87
SKY	39	11	9	21	75

B. Results for Similarity Analysis module

Similarity analysis module was tested against a news sample of 100, containing news stories that are modified by inserting non related sentences into the news story. Then the similarity analysis is carried out and filtered news items were analyzed. Then a percentage of accuracy levels were calculated by using the number of sentences that correctly accepted and rejected for each and every news item in the sample. Then the average of all these calculated percentage values were calculated.

At first, this analysis was carried out without enabling text filters. Then the filters were enabled and test was repeated. So all the texts are filtered before been analyzed for the similarity. This was done manually and it was a bit time consuming and tedious task.

TABLE III
RESULTS OF SIMILARITY ANALYSIS

Similarity Analysis Mode	Accuracy
Similarity Analysis without text filters	69.82%
Similarity Analysis with text filters	82.10%

C. Results for the overall implementation

1) Results for the DOM based Approach: DOM based approach is able to return the main news story of a given page along with the correct image. But sometimes this approach is resulting with the texts that belong to the footer of the page along with the news story. This happens due to the fact that these non relevant texts are accepted by both the filters and the similarity test.

2) Results for the hybrid approach: The above mentioned flaw is overcome in the hybrid approach, as it is capable of extracting main visual area that contains the news story. So the footer notes are not appearing in the news story, resulting with a higher accuracy.

There is a significant performance improvement for the hybrid approach over the DOM based approach. This is due to the fact that only the main visual area is processed by the DOM based approach rather than the whole page.

D. Results for clustering news items

In order to cluster news stories, a threshold value is used. Threshold values between two similar news items are identified as 0.35 to 0.4. For the first time 85 news items were clustered and 69 clustered were obtained with the .38 threshold value. All the clustered news stories had a relationship when considering the meaning of them.

VII. CONCLUSION

Existing methods for extracting information based on the layout of the corresponding web page do have their own advantages and disadvantages. It is obvious that depending on a single approach for extracting information based on the layout is not effective when compared to the hybrid approaches. The combination of the proximity based and the DOM based approaches results with the higher accuracy and increased performance.

ACKNOWLEDGMENT

Mr. Tharindu Dissanayake of Creative Solutions Limited helped us immensely as the industrial mentor to make this project a success.

REFERENCES

[1] Burget, Radek. Layout Based Information Extraction from HTML Documents. In Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference, Parana. Vol. 2. 2007.

[2] S. Gupta, G. Kaiser, D. Neistadt and P. Grimm, Dom-based content extraction of html documents. In WWW2003 proceedings of the 12th Web Conference, pages 207-214, 2003.

[3] T. W. Hong and K. L. Clark. Using grammatical inference to automate information extraction from the Web. Lecture Notes in Computer Science, 2168:216+, 2001.

[4] X. D. Gu, J. Chen, W. Y. Ma, and G. L. Chen. Visual based content understanding towards web adaptation. In Proc. Adaptive Hypermedia and Adaptive Web-Based Systems, pages 164–173, 2002.

[5] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. VIPS: a Visionbased Page Segmentation Algorithm. Microsoft Research, 2003.

[6] Deepayan Chakrabarti, Ravi Kumar, Kunal Punera. "A Graph-Theoretic Approach to Webpage Segmentation", 2008 [online], Available: http://www2008.org/papers/pdf/p377-kumarA.pdf. [Accessed 2008 August]

[7] "Jericho HTML Parser", [Online], Available: http://jerichohtml.sourceforge.net/doc/index.html [Accessed: Aug. 10, 2008]

[8] Simon White, "Tame the Beast by Matching Similar Strings" March 2005. [Online]. Available: http://www.catalysoft.com/articles/MatchingSimilarStrings.html?article=Tame_the_Beast_by_Matching_Similar_Strings_14 [Accessed: Aug. 13, 2008].

[9] Simon White, "How to Strike a Match" March 2005.[Online]. Available: http://www.catalysoft.com/articles/MatchingSimilarStrings.html?article=Tame_the_Beast_by_Matching_Similar_Strings_14 [Accessed: Aug. 13, 2008].

[10] Stephen P. Borgatti, "How to Explain Hierarchical Clustering", 2004 [online], Available: http://www.analytictech.com/networks/hiclus.htm, [Accessed 2008 December]

[11]2008 [online], Available http://alias-i.com/lingpipe/demos/tutorial/cluster/read-me.html, [Accessed 2008 December]

[12]2008 [online], Available http://alias-i.com/lingpipe/docs/api/com/aliasi/classify/PrecisionRecallEvaluation.html, [Accessed 2008 December]