Semantic Based Text Block Segmentation Using WordNet

Nyein Myint Myint Aung and Su Su Maung

***Abstract*—Text block segmentation plays an important role in image search systems. Web pages are segmented into blocks for getting text around the image, which will later be used in retrieval process based on user query keywords. This paper presents semantic based text block segmentation for image retrieval system. Text block segmentation is performed for image indexing using semantic relevance between text blocks. Before segmentation process, web page is broken down in DOM (document object model) tree structure, where tags are used as nodes in the tree. Relatedness between text blocks are computed based on semantic relevance between terms. Semantic relevance in this paper is computed based on the hierarchical structure of the words and their hierarchical information is obtained from Wordnet dictionary. Image search by semantic relevance hence improves the accuracy of the search process**

***Index Terms*—DOM tree, semantic relevance, text block segmentation.**

I. Introduction

With the explosive growth of both World Wide Web and the number of digital images, there is more and more urgent need for effective Web image retrieval systems. To search for images, a user may provide query terms such as keyword, image file/link, or click on some image, and the system will return images "similar" to the query [1].

Most of the text-based mage retrieval system, the images are first annotated by text. And then text-based Database Management Systems are used to index and retrieve WWW images [2]. Above these traditional image management systems, manual image annotation is too laborious and even impossible. To overcome this difficulty, most of commercial web image search systems, such as Google, Lycos, and AltaVista use surrounding blocks of text to index the corresponding images. However, no standard work exists as how to correlate text blocks to web images. Most works use HTML document content structures, such as image file name, page title, ALT-tag, some form of surrounding text and etc. The first three features do not give sufficient information on an image. Among then, surrounding texts are important to index the corresponding Web image. Using relevant surrounding text with respect to an image in an HTML document is different from work to work. Most previous works for web image searches make use of the first paragraph which contains web images as the associated text to index the corresponding web images [3]. In many cases, the first paragraphs containing web images may not have enough text to represent the semantics of web images, thus cause the text-based index with lower performance to support text-based query. However, if more texts surrounding web images are used to index corresponding web images, additional irrelevant words may reduce the performance too. Web page segmentation plays an important role in image search systems. Poor web page segmentation system leads to poor accuracy in image search results. Correct web page segmentation improves the retrieval of the data from the pertinent segments in the page. This paper presents text block segmentation for image search systems by computing semantic relevance between terms. Semantic relevance is computed based on hierarchies of terms which are extracted from WordNet dictionary. The remainder of the paper is organized as the following. Section II is related work. In addition, Section III is text block segmentation and Section IV represents vector space model. In Section V reports the experimental results. Section VI is the conclusion of the system.

Manuscript received January 26, 2013; revised May 10, 2013.

The authors are with University of Technology, Yatanarpon Cyber City,Pyin O Lwin, Myanmar (e-mail: thae.thae.star@gmail.com, susuela@gmail.com).

II. Related Work

The retrieval of images from the Web has received a lot of attention recently. Most early systems employ an essentially text-based approach that exploits how images are structured in Web documents. Sanderson and Dunlop [4] were among the first to model image contents using a combination of texts from associated HTML pages and pages that are one to several links away. They modeled the contents as a bag of words without any structure.

There are some works on using the text for Web image indexing [5]. Their common weak points in processing associated text can be categorized into two types: either only the small parts of the associated texts are selected for the index, or the large associated texts are not elaborately partitioned. The first weakness may cause lower recall and the second one may decrease the precisions of the retrievals.

Many text segmentation methods by topics have been proposed recently. Usually, they obtain linear segmentations, where the output is a document divided into sequences of adjacent segments [6], [7]. Another approach is a hierarchical segmentation; the outputs of these methods try to identify the document structure, usually chapters and multiple levels of sub-chapters [8].

The web page segmentation has been explored by various researchers. There exist various approaches to segment a web page. The Vision based Page segmentation (VIPS) process explained by [9] provides the segmentation roach based on visual features. Since this approach is closer to the human perusal of a web page, the segmentation for this work has been carried out with this approach.

Similarity is a complex concept which has been widely discussed in the linguistic, philosophical, and information theory communities. Semantic typing can be classified into

DOI: 10.7763/IJCCE.2013.V2.257

terms of two mechanisms: the detection of similarities and differences.

Vector space model(VSM) is a popular model for document representation in document and text processing. Documents are represented by vectors of weights, where each weight in a vector denotes importance of a term in the document. In the standard VSM, however, semantic relations between terms are not taken into account. Two terms with a close semantic relation and two other terms with no semantic relation are both treated in the same way. This unconcern about semantics could reduce quality of the segmentation result. In this paper, semantic relatedness between terms are computed in VSM model to get the semantic relation between two text blocks.

III. Text Block Segmentation

Text block segmentation is the basic process of web browsing and image search system. It breaks a large page into smaller blocks, in which contents with coherent semantics are keeping together. In image search system, texts around the image are considered to be relevant text of that image. For the text blocks without image, it is necessary to merge with text block with images.

Any web page can be represented as a DOM tree, with nodes as HTML tags. It is supposed that text *tb* which is contained in the same HTML tag element with image *i* is more relevant to *i*. However, for the relevance of text *tb* if it is not in the same tag element with the image, semantic cohesion is necessary to compute. In this paper, relevant text block of web image *i* is recursively computed to include all the siblings of *tb* if their semantic cohesion is higher than a given threshold.

In order to measure the semantic cohesion of text blocks, semantic relevance between terms are computed. All the child nodes under a parent node are merged bottom up recursively until there is no words similarity between those child nodes or another web image is found under this node. Then, web page will be partitioned into blocks and the image is indexed using the terms of text block which contains that image.

There are two main processes in the text block segmentation process. In the web page segmentation process, firstly, DOM tree is constructed from collected web pages. Then text blocks are segmented according to DOM tree by using semantic relevance algorithm.

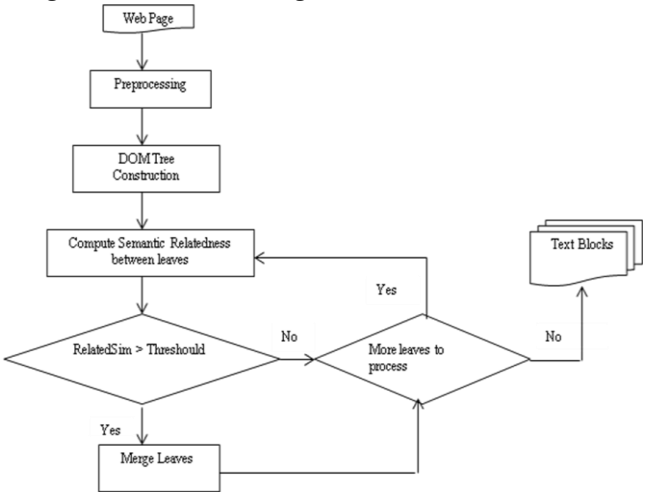

Fig. 1. Overview processes of text block segmentation

Algorithm: TextBlockSegmentation
Input: Webpage W, Threshold t
Output: TextBlock []tb
Begin
DomTree T = BuildDomTree(W)
Flag = true;
For (int i = 1; i < T.leaves.count – 1; i++)
Leave l1 = T.leaves[i];
Leave l2 = T.leaves[i+1];
If l1.contains(image) && l2.contains(image)
continue;
Sim = ComputeSimilarity(l1, l2)
If (sim > t) Merge (l1, l2)
End for
tb = T.leaves
End

Fig. 2. Text block segmentation algorithm

A. DOM Tree Constructions

For the web page segmentation, web page is represented as a DOM tree, with nodes as HTML tags. The HTML DOM views a HTML document as a node-tree. All the nodes in the tree have relationships to each other. The HTML DOM views a HTML document as a tree- structure. The tree structure is called a node-tree. All nodes can be accessed through the tree. Their contents can be modified or deleted, and new elements can be created. The node tree below shows the set of nodes, and the connections between them. The tree starts at the root node and branches out to the text nodes at the lowest level of the tree. Dom Tree is constructed as follows:

- Some of the nodes such as <script>, <font> and comments <!-- --> are not processed. So they are removed in cleaning process.
- Tags such as <IMG> and <A> cannot be parent node, they can only be leaf node (no children) in building DOM tree.
- Each element has exactly one parent node.

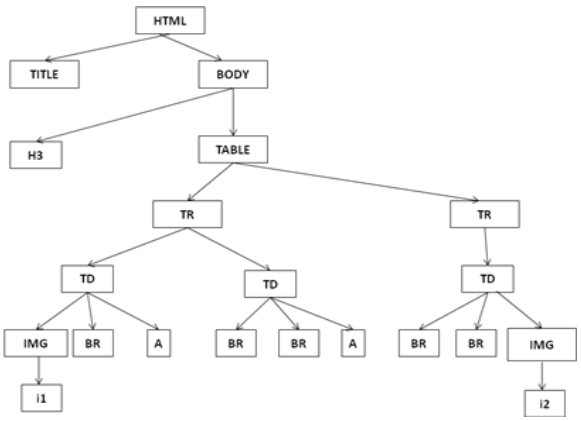

Fig. 3. Example of DOM tree

Text block segation in this paper uses the tag informentmation. the segmentation is then based on the type of the tags. Useful tags include <P> (paragraph), <TABLE> (table), <UL> (list), <H1>~<H6> (heading), etc.

B. Semantic Relevance

Merge all the child nodes under a parent node bottom up

recursively until there is no words similarity between those child nodes or another web image is found under this node. Then, web page will be partitioned into blocks and the image is indexed using the terms of text block which contains that image. In this paper, semantic similarity between terms is computed in order to get overall semantic relevance between text blocks. Vectors are prepared for texts within blocks as in vector space model. Semantic similarity between terms is computed as follows.

The semantic similarity as a function of the distance between two terms in the WordNet hierarchy using edge-based method. Suppose t_1 and t_2 are two terms, and t is their lowest common ancestor. The distance method [10] counts the number of edges connecting the root with t, and edges connecting t with t_1 and t_2. The distance between t_1 and t_2 is calculated us Eq. (1) below and can be easily converted to a similarity value:

$$dist(t_1, t_2) = \frac{len(root, t)}{len(root, t) + len(t, t_1) + len(t, t_2)} \quad (1)$$

where $len(x, y)$ is the length of the path between the node x and y, represented by the number of edges on the path. The distance method assumes that the weight of each edge is always 1. When the parent of the most common ancestor t is the root node, the length of the path between *root* and t is set to 1. It scores between 1 (for similar concepts) to 0. Root word of input terms is found by using WordNet dictionary.

C. WordNet Dictionary

Wordnet is large scale semantic lexicon for the English language. It was started in 1990 as a language project by George Miller and Christiane Fellbaum at Princeton. As of 2006, the database contains about 150,000 words organized in over 115,000 synsets for a total of 207,000 word-sense pairs.

Glosses: computer (a machine for performing calculations automatically)

Links between derivationally related noun/verb pairs: computer, computing, computed, etc.

IV. Vector Space Model

Vector space model is widely used model in information retrieval system. The basic idea is to compute a measure of similarity between query and each document. Queries and documents are represented as vectors in a n-dimensional space (where n is the number of indexing terms) and then compared applying a measure of similarity such as Euclidean distance or the cosine of the angle between the query and document vectors.

Query: Q = (q_1, ..., q_i, ..., q_n)
Document: D = (d_1, ..., d_i, ..., d_n)

Vector space model begins with text documents. In order for each text document to be represented mathematically, each document is turned into a vector. A particular term associated with a given document is represented by a component in the vector for that document. Then, a database containing d documents and t terms is represented by a t x d term-by-document matrix A. The d vectors representing the d documents are the columns of matrix A. The a_{ij} component of matrix A reflects the weighted frequency of the i^{th} term associated with the j^{th} document. Thus, the columns of matrix A are the document vectors and the rows of matrix A are the term vectors.

Fig. 4. Vector Space Model

Weights in Fig. 3 are filled with semantic similarity values computed in above section.

A. Cosine Similarity Algorithm

The cosine of the angle between two vectors is most commonly used as a measure of similarity between documents or relevance of a document for a specific query. Textual similarity is used in the information retrieval community to measure the relevant degree between the document and the information of the user need. Each document and the query are treated as an n-dimensional vector $<w_1, w_2, ..., w_n>$, where n is the number of unique terms in the vocabulary set, and w_i is the term-weight in the document (or the query) concerning in the i^{th} word in the vocabulary set.

To measure a similarity between each document and the query, the cosine coefficient can be computed between the document vector and the query vector; Cosine similarity algorithm is as follows:

$$Sim(X1, X2) = \frac{\sum(tx1_j, tx2_j)}{\left(\sum(t^2 x1j + \sum(t^2 x2j))\right)^{\frac{1}{2}}} \quad (2)$$

Advantages of Cosine Similarity Algorithm:

- Simple, mathematically based approach
- Provides partial matching and ranked results
- Works well in practice
- Calculate the similarity values in a precise way and express the differences in terms of bits of information.
- Allows efficient implementation for large document collections

V. Experiments

For the image search system, web page segmentation (text block segmentation) plays an important role since text relevant to image is an important factor for user keyword. It is necessary to compute semantic relatedness between text blocks. Relatedness between text blocks are computed based on semantic relevance between terms. Terms in text blocks are prepared as vector space model. Weights in vector space model are filled with similarity weight.

Two standards recall and precision, classically used in information retrieval, are employed to evaluate page segmentation algorithms. In the context of web page segmentation, precision is defined as Equation 3:

$$P = \frac{\text{Number of correctly segments}}{\text{Total number of output segments}} \quad (3)$$

Recall is defined as Equation (4).

$$P = \frac{\text{Number of correctly segments}}{\text{Total number of correct segments}} \quad (4)$$

In this evaluation, 300 web pages are used in the experiments with total 2450 segments. Different threshold values are set for experimental purpose in computing semantic relevance between text blocks. Table I shows experiments for different thresholds.

TABLE I: EXPERIMENTAL RESULTS FOR DIFFERENT THRESHOLD

No.	Threshold	Precision	Recall
1	0.1	82.307%	87.34%
2	0.2	88%	89.79%
3	0.3	95.83%	93.87%

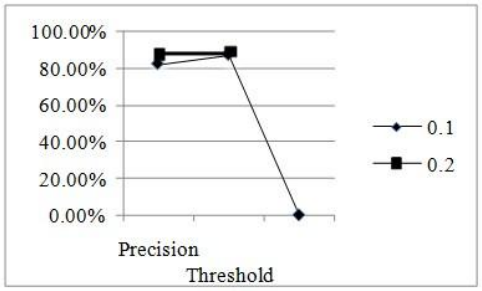

Fig. 5. Performances of Different Thresholds

VI. CONCLUSION

This paper presents the process of text block segmentation for image search system. Web pages are broken down into DOM tree structure based on its tag information. Then it partitions web pages into several text blocks based on their semantic cohesions, blocks which contain web images as associated texts for the corresponding web images. In the text block segmentation, semantic text similarity is computed using similarity by path length. WordNet dictionary is used to get semantic relation between terms. Finding semantic relation in text block segmentation improves relevance level of keyword search since texts around image plays an important factor in image search system.

REFERENCES

[1] *Image Retrieval, From Wikipedia*, the free encyclopedia.
[2] B. Luo, X. Wang, and X. Tang, "A World Wide Web Based Image Search Engine Using Text and Image Content Features," Department of Information Engineering, The Chinese University of Hong Kong, 2003.
[3] H. T. Shen, B. C. Ooi, and K. L. Tan, "Giving meanings to WWW images," in *Proc. Eighth ACM Int'l. Conf. Multimedia*, pp. 39-47.
[4] H. M. Sanderson and M. D. Dunlop, "Image retrieval by hypertext links," ACM SIGIR, pp. 296-303, 1997
[5] Z. Chen, W. Liu, and F. Zhang, "Web mining for Web image retrieval," *Journal of the American Society for Information Science and Technology*, vol. 52, no. 10, pp. 831-839, 2001
[6] M. Hearst, "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages," *Computational Linguistics*, vol. 23, no. 1, 1997.
[7] L. Hernández and J. Medina, "TextLec: A Novel Method of Segmentation by Topic Using Lower Windows and Lexical Cohesion," in *CIARP*, pp. 724–733, 2007.
[8] R. A. Perez and J. E. M. Pagola, "An Incremental Text Segmentation by Clustering Cohesion," Advanced Technologies Application Centre, 2010.
[9] D. Cai, S. Yu, J. Wen, and W. Y. Ma, "A Vision-based Page Segmentation Algorithm," Tech. Rep. MSR-TR-2003-79, 2003.
[10] V. Pekar and S. Staab, "Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision," in *Proc. Int. Conf. on Computational Linguistics*, 2002.

Nyein Myint Myint Aung received the Bachelor of Computer Science from University of Computer Studies, Yangon in 2006 and Master from the University of Computer Studies, Kyaing Tong in 2009. She is working as Tutor in Computer University, Pyay since 2010. Her current research interest includes information retrieval and image processing.