Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 9 (2017) pp. 2887-2895

http://www.ripublication.com

User query based web content collaboration

Dr. Neetu Narwal

Assistant Professor, Maharaja Surajmal Institute,
New Delhi, India.

Abstract

The research aims at providing a tool and methods to extract blocks of information from different related websites and integrate them to provide web user with content specific custom web page. The presently existing tools rely on manual clipping of pieces of web page and integrating them in a single web page. In this paper a comprehensive approach is described that extract page blocks by matching the query entered by the user and automatically suggesting them the related blocks from the web page that can further be integrated in a customized web page.

Keywords: Information Extraction, Cosine Similarity, Web Information Retrival, Page Segmentation, Visual Blocks, Embedded Objects.

INTRODUCTION

Availability of enormous data related to different domain on Internet has resulted in unprecedented opportunities in data-driven knowledge acquisition and decision making. The effective use of available resources poses several challenges: a) Data reserves are plenty, ever changing and distributed. However, it is not possible to collect related data at one place or centralized server. Hence, there exists a need of tool/service that can gather related data from distributed source. b) Internet is not owned by single organization or group, hence the ownership reside with the developer. Therefore, it is impossible that varied resource is available in similar structure or standard. c) The type of operations that can be performed on the data and allowed mode of interaction with the data is quite diverse. Hence it needs formulating strategies for obtaining the necessary information within imposed constraints. d) Data sources available on internet varies in terms of structure i.e., tabular data, relational data, text file and content. Each website uses their own set of concepts, attributes and

relations to represent the data. Hence effective data integration strategy from different sources needs to be discovered.

This paper presents a data integrating and collaborating tools that gathers varied data from different data source and integrates them based on the user needs. Tool is equipped with machine learning algorithm for web page partitioning and similarity based matching algorithm is used for data clipping and further integration.

The rest of the paper is organized as follows. Section 2, presents the related research work done in the area of web content integration. Section 3, describes the methodology used for user query based web content collaboration. Experimental evaluation is presented in Section 4 to find the feasibility of the tool.

RELATED RESEARCH WORK

Information Extraction means extracting structured information from the available pool of data i.e., news corpora, medical corpora, manuals, data dictionary etc. Lately Information Extraction has started moving towards the web as source of textual information. Thus giving rise to new research area namely Web Information Extraction. The web data is highly unstructured data and thus it poses difficulties to directly implement information extraction techniques on the web. Unstructured nature of the web poses certain challenges like authentication, authorization, validity of data.

Researchers have suggested the database techniques for building specialized program called wrapper for web information extraction, which maps the data of interest into specific format. A wrapper is a specially designed procedure for extracting content from information source and delivers the content of interest in a descriptive representation. A wrapper for a web source accepts queries about information and returns the result[3]. It consists of a set of extraction rules and the code required to apply these rules and is specific to one source. To extract information from several independent sources, the libraries of wrappers are needed. Wrappers should be able to cope with the changing and unstable nature of the web.

The construction of a wrapper can be done manually, or by using a semi-automatic or automatic approach. The manual generation of a wrapper often involves the writing of static code. The developer understands the document structure and then translates it into program code. One of the first approaches to the framework for manual building of web wrappers is the Stanford IBM Manager of Multiple Information Sources (TSIMMIS) system. The goal of TSIMMIS system is to provide tools for accessing in an integrated fashion, multiple information sources, and to ensure that the information obtained is consistent.

Semi-automatic wrapper generation benefits from support tools to design the wrapper. Some approaches offer a demonstration-oriented interface where the user shows the

system what information to extract.

Automatic wrapper generation uses machine-learning techniques, and the wrapper research community has developed learning algorithms for a spectrum of wrappers from the very simple to the relatively complex.

ShopBot [2] is a comparison shopping agent, specialized to extract information from web vendors. The algorithm focuses on vendor sites with form-based search pages returning list of products with a tabular format. Information is extracted from resulting pages using a combination of heuristic search, pattern matching and inductive learning techniques.

The Wrapper Induction Environment [5] is a tool for assisting with wrapper construction. It works on structured text containing tabular information, and it is demonstrated for HTML documents. Soft mealy [4] is a system that learns to extract data from semi-structured web pages by learning wrappers specified as non-deterministic finite automata. An inductive generalization algorithm is used to induce contextual rules from training examples.

Recently, there have been several tools developed to integrate web page contents. Jie et. al. [7] designed a tool namely Homepage Live that pick content from DIV tag from different website. These contents are then manually combined to provide integrated web content to the user.

Kowalkiewicz et al. [8] provides a tool called MyPortal to select data blocks, such as paragraphs, tables, lists, and then integrate them automatically using IFrames. Using the tool the CSS formatting of the original web data is lost, and so users receive differently formatted results.

ClipMark (Clipmarks) is a web browser plug-in to extract fragments of static web pages. However, the fragments are not updated automatically when the original web page changes.

WEB CONTENT COLLABOARATION SYSTEM

This system presents a web content collaboration tool that gathers information from different website returned from Google search, and dynamically pick the content related to the query entered by the user. The related content are gathered and collaborated into a final web page that shows user perceived content from non-related websites.

The system comprises of the following modules

- Web Page Gathering Module
- Web Page Segmentation Module

- Content Filtering
- Content Collaboration and Rearrangement

The first module accepts the query from the user and passes the query to the Google Search engine and pick top 10 web sites returned as a result for processing in the next module.

These listed web pages are uploaded and passed to the next module for segmentation, this module accept the web page as input and parse the tree structure of the web page using the approach of top down parsing and utilizing Document Object Model API functions. The input web page is recursively traversed and broken down into blocks based on the heuristic rules [1]. The semantically cohesive content comprises of the visual block if the block size reaches below a lower threshold then it is merged with the sibling nodes to obtain visual blocks.

The obtained visual blocks of each page is then compared with the query terms to identify the visual blocks that are having similarity to the user query. The comparison is done based on cosine similarity measure. For each visual block a vector is prepared comprising of terms in block. The system then estimates the similarity of term vector with the query terms.

The cosine similarity measure is computed as the cosine of the angle between the term vector dj and query vector q as mentioned in Equation 1.

$$Cosine(d_j, q) = \frac{<d_j.q>}{|d_j| \times |q|} \qquad (1)$$

$$= \frac{\sum_{i=1}^{|V|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{|V|} w_{ij}2} \times \sqrt{\sum_{i=1}^{|V|} w_{iq}2}}$$

where w_{ij} = term weight in block

w_{iq}=term weight in the query it is constant value 1.

In the next module the highest matching visual block is extracted from the top 10 web sites. These blocks are stored in an XML file storing the features related to height, width and other formatting features of these blocks. Finally a new web page is designed by picking the content from the XML file and arranged according to the device screen dimensions.

The final output is the website comprising of only relevant information from top 10 related websites listed by search engine. The website is free from noise content which occupy the major portion of these websites, thereby saving time and effort required to reach to the related content.

The final rearrangement module checks the browser screen dimension for creating the new webpage. Hence this tool can be utilized for different devices types like mobile phone, palm tops, Personal Computer, Laptops etc. for providing better user experience.

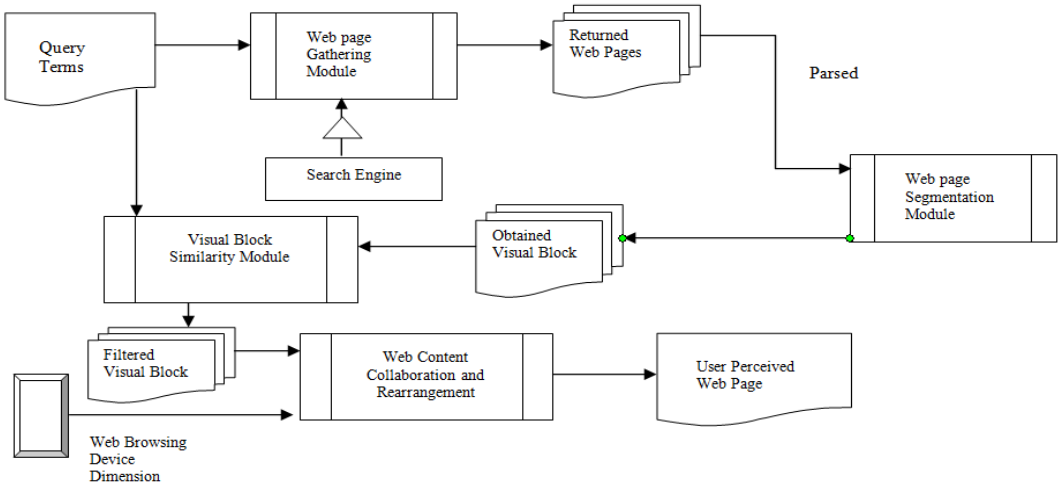

Figure 1: Methodology used in the Research Work

EXPERIMENT AND RESULTS

Experiment is conducted with top 15 news web sites returned by Google Search Engine "Paneerselvam Sasikala Controversy" as on 13 Feb 2017. These web pages are parsed with segmentation algorithm and obtain 250 visual blocks extracted from these web pages.

To derive the identification of each block, we have used the approach of learning by example, where the dataset is manually pre labeled with class and trained to build a model. Each block is represented as (x, y), where x is set of similarity measure of each block and y is the class. In this paper Artificial Neural Network (ANN) techniques to train the model. For each block cosine similarity measure is computed between the query and the visual block text content. The result obtained is used for manually labeling the blocks as Match or No Match.

The dataset is divided into 80-20 ratio where 80% records comprise of training records and 20% comprise of testing records. The system is trained with feed forward Artificial Neural Network classifier. The model is tested for rest 20% of records and the result is evaluated using different evaluation measures such as Accuracy, Precision, Recall, F-Measure.

The classifier predictive capability evaluation measure is computed using confusion matrix as shown in Table I. The confusion matrix is the table of size m by m where m is the total number of class in the dataset, where each row depicts the actual outcome or class given by the classifier and each column depict predicted outcome or class. True Positive (TP) and True Negative (TN) are indicators of correctness of the classifier. Whereas, True Negative (TN) and False Positive (FP) are indicators of error or mislabeled tuples [11].

Table 1: Confusion Matrix

		Predicted Class		**Total**
		Yes	No	First
Actual Class	Yes	TP	FN	P
	No	FP	TN	N
	Total	P’	N’	P+N

The accuracy of classifier is the percentage of test tuples that are correctly classified by the classifier.

$$Accuracy = \frac{(TP+TN)}{(P+N)} \quad (1)$$

Precision is a measure of exactness means percentage of tuples labeled as positive.

$$Precision = \frac{(TP)}{(TP+FP)} \quad (2)$$

Recall is a measure of completeness means percentage of positive tuples labeled as positive.

$$Recall = \frac{(TP)}{(TP+FN)} \quad (3)$$

F-measure is a combination of precision and recall.

$$F - Measure = \frac{(2\ x\ Precision\ x\ Recall)}{(Precision+Recall)} \quad (4)$$

We have used feed forward neural network, where the input layer has one neuron and output layer has two neurons. The sigmoid activation function is used to train the model and performance is evaluated after performing five-fold cross validation. Table II shows the efficiency of the classifier depicted in terms of evaluation measures.

Table 2: Accuracy Measure of Classifier

Feature set	Accuracy	Precision	Recall	F-Measure
Feed Forward Neural Network	0.9603	0.8832	0.9295	0.9056

In order to test the content collaboration system, the tool is tested with few popular newspaper websites and the topic search was "Paneerselvam Sasikala Controversy" as on 13 Feb 2017. The Top 10 websites are parsed with segmentation algorithm and obtain 82 web page blocks extracted from the web pages. Each block is matched with the query terms using cosine similarity measure. The highest related visual blocks from each website is returned and stored in a XML file. Finally the blocks are rearranged according to browser screen dimension.

Figure 2 displays the top 3 websites showing news related to the query entered by the user. As it can be observed that the related content block occupies a small portion of the web page, the whole page comprises of other non related news, advertisement, menu items, hyperlinks etc. Therefore the main essence of searching for the topic is lost sometimes.

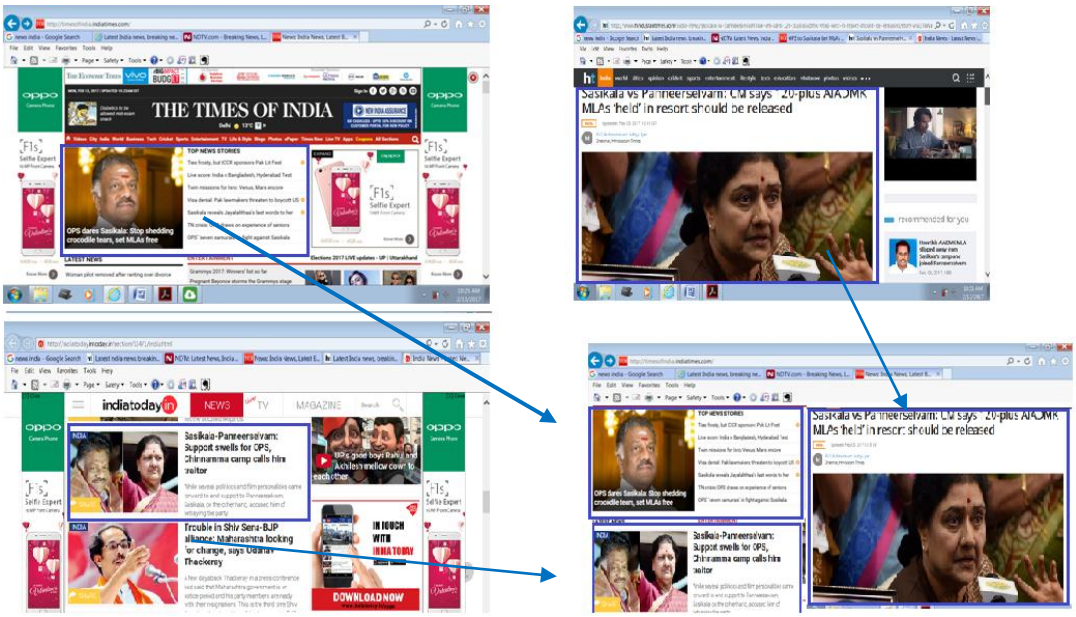

Fig 2: News related to "Paneerselvam Sasikala Controversy" highlighted in the four popular news websites

The content collaboration system combines news from these websites into a single page without noise content. The new web page generated has the pure news and it is of content to the user. The web page shown in the right bottom is the final output page which gathers the content from different web pages and produces an output which is clean, free from noise and purely related content based on user query. The system extracts the content and collaborate them in the new web page.

CONCLUSION

We have presented an approach to building a system that pick and integrate contents from different websites based on user query. The system demonstrates the semantic web concept and provides user perceived content gathering and collaboration. The contribution of the work is in presenting a system enabling new methods of browsing the Internet, clipping web pages to improve relevance of browsed content and to improve user experience in mobile, small screen, devices. We have also shown ways of integrating the content that may facilitate business use of the Internet. Preliminary research results indicate that the system is more robust than current approaches in Web data collaboration.

REFERENCES

[1] N Narwal, S K Sharma, Amit Prakash Singh, Entropy based content filtering for Mobile Web Page Adaptation, Proceeding WCI '15 Proceedings of the Third International Symposium on Women in Computing and Informatics Pages 588-594 , ACM New York, NY, USA 2015.

[2] Doorenbos R.B. , Etzioni O., Weld D.S., A Scalable Comparison-Shopping Agent for the World Wide Web. Technical report UW-CSE-96-01-03, University of Washington, 1996.

[3] Eikvil , Information Extraction from World Wide Web - A Survey - Technical Report 945, Norvegian Computing Center, 1999.

[4] Hsu C. H. and Dung M. T., Generating Finite State Transducers for semistructured Data Extraction from the web. Information Systems, Vol 23, No. 8, pp 521-538, 1998.

[5] Kushmerick N., Weld D.S., Doorenbos R., Wrapper Induction for Information Extraction. Ph.D. Dissertation, University of Washington. Technical Report UW-CSE-97-11-04, 1997.

[6] Similarity-based web clip matching, Małgorzata Baczkiewicz, Danuta Łuczak and Maciej Zakrzewicz, Control and Cybernetics, vol. 40, Issue 3, 2011.

[7] Jie, H., Dingyi, H., Chenxi, L., Hua-Jun, Z., Zheng, C. and Yong, Y. (2007) Homepage live: automatic block tracing for web personalization.16th International World Wide Web Conference (WWW2007). ACM Press,pp- 1–10.

[8] Kowalkiewicz, M., Orlowska, M.E., Kaczmarek, T. and Abramowicz, W. (2006) Towards More Personalized Web: Extraction and Integration of Dynamic Content from the Web. Proc. of APWeb Conference. Springer, 668–679.

[9] Kulathuramaiyer, N. (2007) Mashups: Emerging Application Development Paradigm for a Digital Journal. Journal of Universal Science, 13 (4), 531–542.