Unified Motor Vehicle Sales System

M.A.E.Y. Fernando, K.M.D.P.P. Jayathilaka, N.H.K.V. De Silva, E.M.P.S. Dehideniya,
P. Samarasinghe and H. Perera

***Abstract* — In Sri Lanka, current vehicle sales trading methods are diverse and require a lot of effort in finding a preferred vehicle from web advertisements online newspapers and printed newspaper advertisements. Due to popularity of the newspaper classified advertising, many people tend to use it and the classifieds amount in a newspaper has been increasing drastically. Therefore the buyers face many difficulties in finding a specific vehicle of their preference. This project addresses the problem by integrating all vehicle sales trading methods mentioned above. After integration, the output is given to the user by a web application and a desktop application. In order to achieve this, web scraping, image processing, character recognition techniques and Google Custom Search are used. From these technologies, web scraping and image processing are used for online and printed newspapers respectively. Custom web search module is implemented using Google Custom Search for more focused web searching. As for the future improvements, since the application is based only in English, it is preferred to create using Sinhala and Tamil languages. As an innovative addition, mobile application can be implemented for publishing vehicle advertisements in the web application of the implemented website in multiple mobile platforms.**

***Keywords* — Web Scraping, Image Processing, Web Search, Online Advertisements, Optical Character Recognition**

I. INTRODUCTION

The main software product outcome from this research project is called "Unified Motor Vehicle Sales System". It uses hardware, software and various technologies to fulfill the requirement; to unify the vehicle sales scenario by combining web advertisements (online newspaper classifieds, printed newspaper classifieds and web search results.

Printed Newspapers have been the most popular method of choice for the people to pass any message efficiently and easily over a particular geographical location (In this project, the context is Sri Lanka). Although there is a considerable growth in other media such as television, radio, internet, and mobile phones etc. people tend to use these newspapers very effectively to communicate with all around the country not only in the past but also in the present. In Sri Lanka, there are so many newspapers publishing in both weekday and weekend by several newspaper publishers. In those newspapers, people publish thousands of advertisements in vehicle selling category, which are from all over the country. Vehicle advertising through newspaper classifieds give a concise and clear message to the reader about a particular entity and also for the newspaper agency can advertise many items at the same time while the reader has the chance to compare and contrast each item because of the item abundance.

According to the research findings done by Warushamana and Weerawansa[1] despite there are many opinions that people tend to read fewer newspapers, the study has shown that there has been no significant decline in newspaper circulation and there is no threat to from the internet to newspapers, the newspaper circulation has been increased over the past ten years.

M.A.E.Y. Fernando, K.M.D.P.P. Jayathilaka, N.H.K.V. De Silva, E.M.P.S. Dehideniya, P. Samarasinghe, and H. Perera are with Sri Lanka Institute of Information Technology. (e-mail: emmanuelyasith@gmail.com, asan4all@gmail.com, kalanavinura@gmail.com, praveen.dehideniya@gmail.com, pradeepa.s@sliit.lk and hansa.p@sliit.lk)

This fact is further proven in [2] the article where the author mentions that newspaper classifieds will prevail with respect to the vehicle marketing section due to its simplicity and closeness to the buyers and sellers.

In Sri Lanka, web advertising has also come a long way since its inception, but still, website advertising in Sri Lanka is still in its growing stage. These web advertising is mostly used by the people who are living in rural areas. So, newspaper classifieds are still dominating this advertising area, even though the web advertising is in its developing stage.

II. RESEARCH METHODOLOGY

The research was carried out mainly for the following domains.

- Web scraping – (For online newspapers)
- Web search – (For Google Custom Search™)
- Image processing – (For extracting printed newspaper classifieds)
- Character recognition – (For identifying characters in the extracted newspaper classifieds)

Architecture of the complete system is mentioned below.

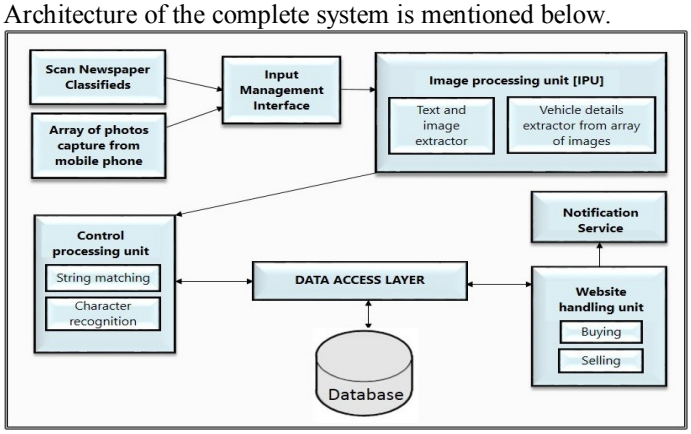

Fig. 1. System architecture

A. Web scraping – (For online newspapers)

Web scraping is the process of extracting and creating a structured representation of data from a website [3]. To extract data behind web pages, an algorithm was built up targeting the markup of the relevant webpage. An XPath expression was evaluated to parse the web page to a tree representation. An XPath denotes a path, possibly with wildcards, and when evaluated on a tree, the result will be the set of nodes with relevant text at the end of any occurrence of the path in the tree. HTML, the markup language used to structure data on web pages, is intended for creating a visually appealing interface for humans [3]. When the XPath expression gets executed, it extracts all the data which is in that particular online newspaper URL. Therefore the extracted data will be processed further by using string matching algorithms to segment the text in order to identify the vehicle manufacturer, model, type, owner, price, address, etc.

B. Web search – (For Google Custom Search™)

In order to provide a more focused web search experience for vehicle seeking users, the web search option was implemented using Google™ search engine which was introduced by S. Brin and L. Page [4]. Since user intentions are subjective and solely based on the opinions. According to Benavides, Caro and Yates [5], other than considering all possible characteristics, a few and more common characteristics were considered. Since the user's intent can be used for efficient web search ranking, according to Agichtein, et. Al [6],[7] & [8]; the customized web search specifically focused but had the flexibility to enter queries based on user preference.

C. Image processing – (For extracting printed newspaper text)

The process of extracting text is a four step process; Image enhancement, segmentation and extracting. Thresholding techniques can be used for image enhancement. Several thresholding methods are available for enhancement are Global thresholding, Otsu's method and local thresholding [9]. Global thresholding is the most method. This is selected because it is most suitable for black and white images.

There are several methods available for image segmentation. Bloomberg's text/image segmentation algorithm is a method to remove images from the newspaper article [10] Segmentation also can be performed using string matching algorithms on newspaper articles. There are single and multiply pattern-matching algorithms. Some are Approximate String Matching Algorithm is based on fuzzy logic, Rabin Karp Algorithm uses hashing techniques and Boyer Moore algorithm [11]. Boyer Moore algorithm uses right to left comparison which is important to our context. Similar method is used segment.

Advertisements are segmented using the unique number provided to each advertisement. Next segmented images are extracted to text using an OCR.

III. RESEARCH FINDINGS/RESULTS AND EVIDENCE

Since the project mainly focused on software implementation based on the research, following are the research findings related to the previously discussed domains.

Among the wide range of software applications being developed today, the Unified Motor Vehicle Sales System software applications are carried out a major role of it. Those software applications are ideal for both business processes and advertising area. Among the people in foreign countries these applications are popular in their daily work, but still there were no such motor vehicle sales systems available in Sri Lanka. By using the implemented system, many people can easily overcome their difficulties. The requirement is to provide efficient, useful and reliable motor vehicle sales advertisements to the users. In this project, the main vision was to research and develop a comprehensive software application that will facilitate people, who are keen in vehicle sales not only for identifying the vehicles which are available for sale and but also to advertise their vehicles in the Sri Lankan market with the correct combination of latest information and communication technologies.

But in other countries has in-built technologies related to this area, because of their highly available resources. And there were many researches for web scraping, newspaper classifieds extraction, Google result filtering and vehicle logo, color, model detection have done on different methods in different countries. But, still there were no any research projects done for motor vehicle sales section in Sri Lanka.

As the outcome of this research-based project, a solution to unify all the vehicle sales advertisements which are published on printed newspaper classifieds and online newspapers to a single location where the people can access these advertisements through implemented system website. Not only searching advertisements, but also this website allows people to publish their vehicle sales advertisements as they expect. So far UMVSS systems have obtained the vehicle sales advertisements which are published on printed newspaper classifieds and online newspapers. Also the system website is capable of filtering Google web results and the previewing of vehicle advertisements obtained from the online newspapers.

Despite having substantial limitations due to the free version of Google™ Custom Search, the user query is handled before it is sent to the Google search engine through in-built functions of the Custom Search API. Since Google search relies on the worldwide web, the web search had to be localized in order to provide more user-friendly results related to Sri Lankan customers. The custom search was also configured to retrieve results based on popular Sri Lankan vehicle sales advertising web sites.

Here are some of the user interfaces of the implemented software product.

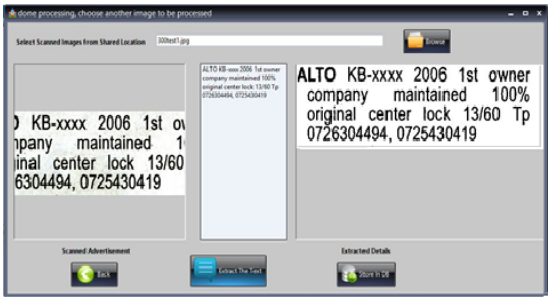

Fig. 2. Newspaper classifieds extractor

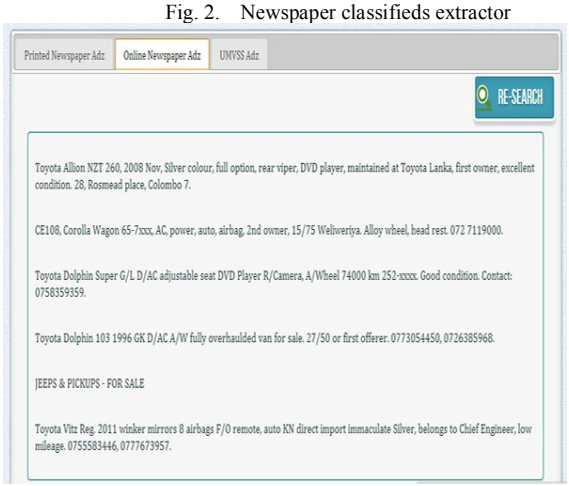

Fig. 3. Printed newspaper ads page

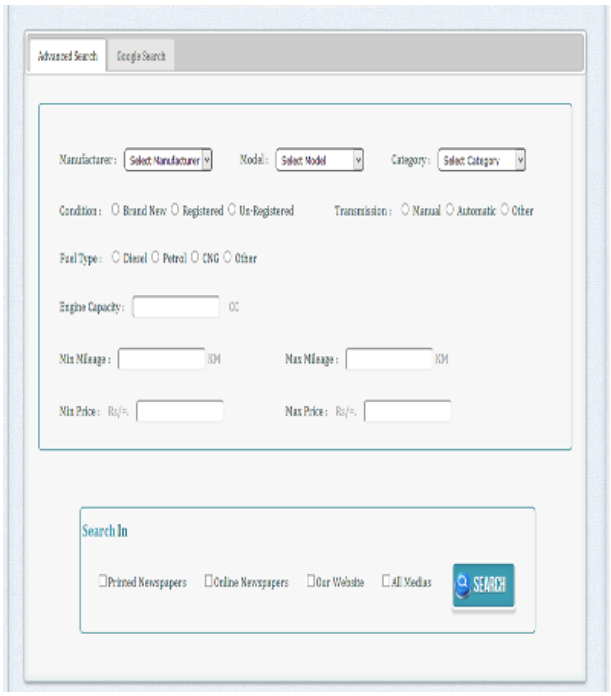

Fig. 4. Advertisement publishing page for UMVSS

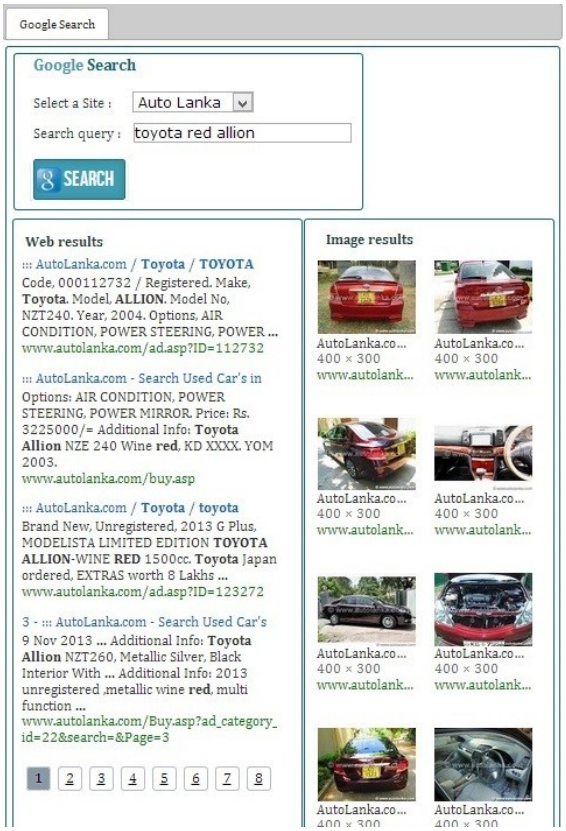

Fig. 5. Google custom search results page

IV. CONCLUSIONS & FUTURE WORKS

In this project the main objective was to unify possible means of vehicle advertising in Sri Lanka. In Sri Lanka, vehicle advertisements are published mainly using printed newspaper classifieds. Considering the web advertising for vehicles in Sri Lanka has many shortcomings such as not having a particular format for advertisements etc.

Future research includes the product implementation to be done in mostly common languages used in Sri Lanka; Sinhala and Tamil. Since both Sinhala and Tamil language support increases incrementally and if there is an accurate OCR for Sinhala and Tamil, it is possible to implement newspaper classified extraction for both Sinhala and Tamil newspaper classifieds, which is convenient for the majority of Sri Lankans. Implementation of a mobile application for mobile application platforms such as Android™, iOS™ and Windows Phone™ is indorsed as future development which enable the users to access the implemented website through smartphones. An artificial intelligence based string manipulation for setting query in for Google search and advanced search functionalities, the accuracy of the web search result set can be improved.

ACKNOWLEDGMENT

The team would like to express the deepest appreciation to all those who provided the possibility to complete this report. The deepest thankfulness goes to project examiner Mr. Indraka Udayakumara, CDAP lecturer-in-charge Mr. Jayantha Lal Amararachchi and all the people who helped to complete this research.

REFERENCES

[1] S.R.S.D.K Weerawansa, G.K Warushamana. "The Traditional Newspaper Industry vs. the Technology Driven News Media - Predatory or Complementary Behavior: Issues, Trends, Prospects and Emerging Realities of the Print Media faced with Emerging Technological Directions in the Sri Lankan Context." [PDF]. Available: www.cmb.ac.lk/academic/arts/Home/title24.pdf [Feb.16, 2013]

[2] S. Diaz. "On the Internet, A Tangled Web Of Classified Ads." Available: www.washingtonpost.com/wpdyn/content/article/2007/08/30/AR2007083002046.hhtml?hpid=sec-tech Aug 31, 2007[Feb. 18, 2013]

[3] P. H. Cording, K. Lyngby (2011) "Algorithms for Web Scraping" [PDF] Available: http://www2.imm.dtu.dk/pubdb/views/publicationdetails.php?id=6183 [Apr. 01, 2013]

[4] S. Brin, L. Page. (1998). The Anatomy of a Large-Scale Hyper textual Web Search Engine [Online]. Available: http://ilpubs .stanford.edu:8090/361/1/1998-8.pdf [Feb. 20, 2013]

[5] L.C. Benavides, C.G. Caro, R.B. Yates. Towards a Deeper Understanding of the Users Query Intent [Online]. Available: http://grupoweb.upf.es/WRG/dctos/Calderon-Gonzalez-Baeza-SIGIR10.pdf [Apr 02, 2013]

[6] E. Agichtein, E. Bell, S. Dumais. Improving Web Search Ranking by Incorporating User Behavior Information [Online].Available:http://web.cs.dal.ca/˜anwar/ir/review/grads.pdf [Apr 20, 2013]

[7] D.I. Hernndez, P Gupta, P. Rosso, M. Rocha. A Simple Model for Classifying Web Queries by User Intent [Online]. Available:http://users.dsic.upv.es/˜prosso/resources/HernandezEtAlCERI12.pdf [May 5, 2013]

[8] B.J. Jansen, D.L. Booth, A. Spink Determining the User Intent of Web Search Engine Queries [Online]. Available:www.www2007.org/posters/poster989.pdf [May 6, 2013]

[9] S. S. Al-amri1, N.V. Kalyankar and S.D. Khamitkar. Image Segmentation by Using Threshold Techniques. [Online] Available: http://arxiv.org/ftp/arxiv/papers/1005/1005.4020.pdf [Mar. 30, 2013]

[10] S. S. Bukharia, F. Shafaitb and T. M. Breuela. Improved Document Image Segmentation Algorithm using Multiresolution Morphology [Online]. Available: http://www.csse.uwa.edu.au/~shafait/papers/Bukhari-Text-Image-Segmentation-DRR11.pdf [Apr. 20, 2013]

[11] N. Singla, D. Garg. String Matching Algorithms and their Applicability in various Applications [Online]. Available: http://www.ijsce.org/attachments/File/Vol-1_Issue-6/F0304111611.pdf [Apr. 30, 2013]