Information Aggregation for Tailored Mobile Information

Mark Oliver, Dr Samuel Mann
School of Information Technology & Electrotechnology
Otago Polytechnic
Dunedin, New Zealand
smann@tekotago.ac.nz

ABSTRACT

Information aggregation is a method of customising content from the Internet for users. There are so many sources of information available to users on the Internet that many people are finding that they don't have to time to visit every site they may want to in the small time that they have to browse in a day. The problem of information availability is exacerbated by user's increasing reliance on WAP/WML technology as a primary information device. Most of the information available on the world wide web is not accessible to such devices. This paper describes a system developed by the authors for providing information aggregation for WAP devices. The paper first describes information aggregation and then two contrasting approaches for identifying and collecting desired information. The process of converting aggregated data from html through XML to WML is also described. The result is an application that will allows users to perform information aggregation on a recurring basis from their personal computer, the product operating on a server and viewable via a WAP phone.

1. INTRODUCTION

This paper describes the development of an application to make information on the web available to mobile users.

1.1 Growth but limitations of mCommerce

Kjell Sorme of Ericcson argues that the "mobile internet is more than making the Internet mobile…it will enable a whole new set of services that can be tailored to the location and individual preferences, needs and circumstances of the mobile user" (Sorme 2001). He describes three phases of mobile Internet development. The first is using the cell phone to provide network connection for laptop computers. The second "more critical wave" is taking the internet to mobile devices. In the mobile phone industry "internet enabled" refers to Wireless Application Protocol (WAP), or 'WAP enabled'. WAP is a communication protocol specifically designed for wireless devices. It allows web pages to be displayed on mobile devices using a WAP browser that uses a version of Wireless Markup Language (WML). The third wave phase is the development of mobile content and applications centered on mobility: situation centric.

Small screens, slow connections, extremely limited processing capability and major concerns over usability (Nielsen 2000) mean it is unlikely that "the web" as it is known to real computer users will be available to the mobile user in the medium term at least. While a derivative of XML in the same way as HTML is, the Wireless Markup Language (WML) used by mobile devices is not HTML. mCommerce requires a separate set of scripts, for organizations struggling with the management of their web presence, the need to duplicate their site is a big ask.

There are some web sites that aim to collect information together that is suitable for WAP. itouch (www.itouch.co.nz) in New Zealand, is a site that collects information suitable for WAP. While there is indeed much information available from such sites, they are not configurable beyond choosing from a range of preselected information, and cannot present the user with the river level on the Upper Taieri or Dunedin club rugby scores. Such sites, collecting a limited subset of information and presenting it in a restricted format break Metcalfe's law that argues against fragmentation of the internet (Nielsen 1999).

The goal of the research described in this paper is to make information in the 'web accessible, and usable, for mobile users.

1.2 Information Aggregation

Information aggregation is becoming a popular way of customising content from the Internet for users. There are so many sources of information available to users on the Internet that many people are finding that they don't have to time to visit every site they may want to in the small time that they have to browse in a day. Many sites only hold pieces of information that a user may want to view on a daily basis, such as the weather and what's on the television. Information aggregation brings selected pieces of information from many different sources on the Internet and collects them in one place, giving users access to selected segments of information that they want to view quickly. This solves the problem of having to visit many different sites in order to gain required knowledge.

Screen scraping is one form of information aggregation. Screen scraping has been used widely in the past for interfacing with legacy systems. Hackathorn (1999) described that in the 1980's there were many applications that behaved like 3270 or vt100 terminals to legacy systems. He argued that "these applications were not reliable, but they were cheap".

More recently screen scraping is being used to obtain information from websites as a method of information aggregation. Vartanian and Ledig (2001) reviewed the development of data aggregation via the 'web. They discuss how data aggregation can become quite formalized in the consolidation of the usernames and passwords (collectively "PINs") that permit them to access a variety of PIN-protected websites that contain information about their personal accounts on a single website by using one master PIN. These online account providers could be financial institutions, stockbrokers, airline frequent flyer and other reward programs, e-mail accounts and any other website offering PIN-protected personal accounts to users. Applications that perform automated collection of information are described as "Web-wrappers" by Sahuguet and Azavant (1999).

By stripping HTML to extract the text, it can easily be reformatted as WML via XML. This means components that cause difficulty for WAP such as images, tables and links are removed. This means much information from multiple sites, or multiple pieces of information from a single site, can be placed onto a series of WML cards. You can get all the information you want onto one WML card form one scrape. If you try to too much information form one scrape then you are defeating the of being able to cut web pages up into little packets of information that you can display on a WA P device.

Information aggregation from HTML and reformatting into WML such that it may be viewed via mobile phones is the approach adopted by this research. This paper describes the development of an application that a home user can operate, that will give them the power to perform information aggregation for themselves. This would allow users to be able to logon to a service provider and perform a series of predefined screen scrapes to quickly gather information from all over the Internet from a number of different sites in a matter of seconds. Scraped information would then be formatted into an WML document which the user would be able to view offline or update a personal

website that a user could view live.

2. METHOD

The goals of the development are to:

- The ability for users to be able to select segments of information from Websites on the Internet and scrape them for personal use.
- The ability for users to be able to customize each information scrape in terms of time occurrence.
- The ability to format scraped information for presentation to users (i.e. present in WML).

The result will be an application that will allow users to perform information aggregation on a recurring basis from their personal computer to be formatted for access via mobile phone. The remainder of this paper describes experimentation in the development of scraping information from remote pages.

2.1 HTML Scraper

The concept is simple. If you can identify unique HTML code above and below the item that you want to extract from an HTML page you could extract the information in between. This approach relies on the html used to generate a page being constant, only the content being dynamic. A prototype was developed using Visual Basic to accomplish the task.

2.2 Process

The Internet Transfer control was used to download the HTML code to the users local computer. Once this had occurred a find method was used to help locate the information that the user intended to scrape within

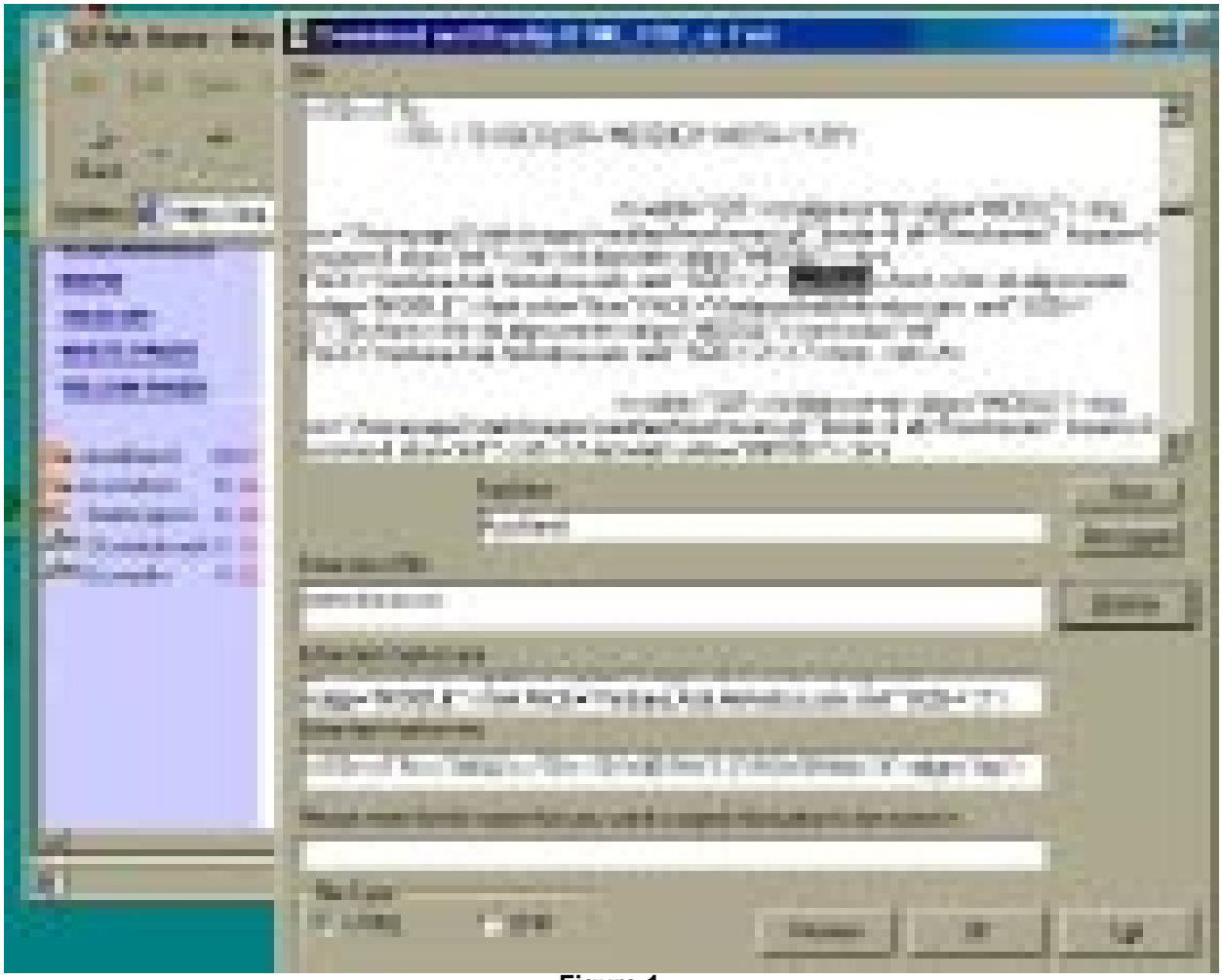

Figure 1

Xtra (www.xtra.co.nz) website on left with scrape application showing method of identifying html markers.

the HTML code of the target URL.

In the example shown in Figure 1 the user is trying to scrape weather information from the Xtra website. They have typed in “Auckland” and located the weather information that they want. Once the start point of the information the user wants to scrape has been located, HTML code is selected from in front of this position. It is then copied and pasted into the text box labelled “Enter text marker one”. The user has to be careful not to select any actual content, has this will change constantly in a dynamic page. If the marker changes that the user has chosen to act as the start or end position of the information that they want, then the scrape will fail. The user must then locate the end of the information that they want to extract and copy and paste some unique HTML code into the text field labelled “Enter text marker two”.

The user can test to see if the HTML code that they are going to use as a marker is unique by using the find method once more. If the HTML code that is intended to be a marker is found more than once in the HTML code, then it is not unique and not suitable to be used as a marker.

Once both HTML markers have been entered the user can preview the HTML code that is going to be extracted by clicking the preview button as shown below the information that will be extracted is highlighted in black (Figure 2).

The user enters a file name that the scraped information for this page is to be stored in and choses either HTML format or Cold fusion format and clicks

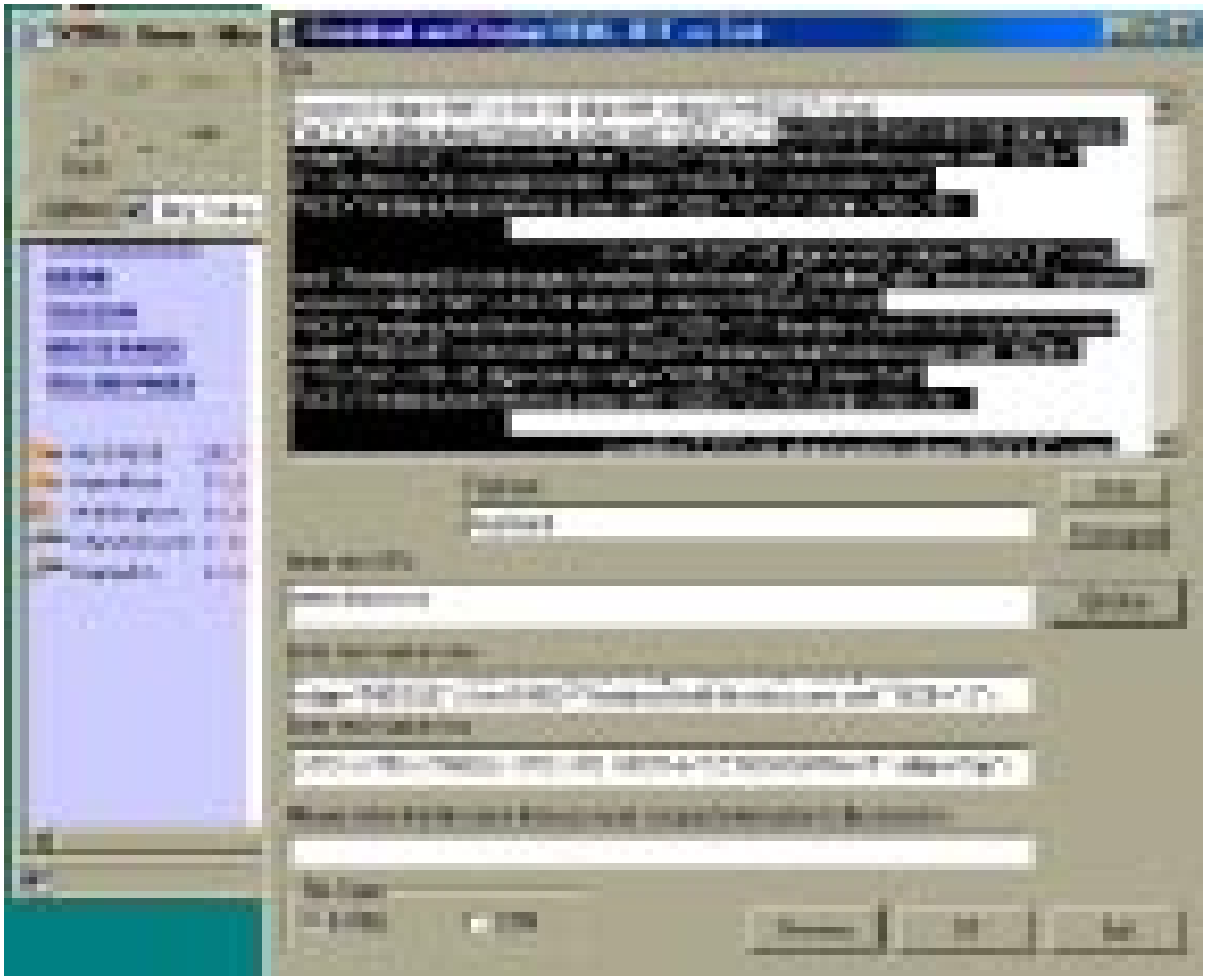

Figure 2

The html and content between the markers is highlighted for confirmation

OK. The user is now able to scrape the information that they have targeted from the URL of their choice. Using the application the user then creates a file that effectively lists the scrapes and schedules their functioning.

This is a cut down list of events that occur for every page that is scraped.

- For each file in the scrape list…
- Information from storage file is extracted into variables that hold:
  - o Page URL.
  - o Start Marker as HTML code.
  - o End Marker as HTML code.
  - o Name of new file that holds the scraped information.
- Microsoft Internet control is used to download HTML content from target page to local machine
- A search for the Start Marker is made in the URLs HTML source code, when found a start location in the file is stored
- A search for the End Marker is made in the URLs HTML source code, when found an end location in HTML source code is stored
- Calculation is made; all HTML code between the start location in the file and the end location in the file is extracted
- The extracted information is stored in the file that the user specified i.e. "Weather.html" (or .wml but is no checking that doesn't contain none WML compliant tags)
- This can then be called remotely

3.2 Issues Involved with this Method of HTML Extraction

There are a number issues that have to be addressed in order for this scraping application to be developed further.

- Another method has to be developed to allow the user to select information that they would like to extract from a Web page. This application does not make it easy for the user to identify what information they want to extract. Also the user has to have an understanding of HTML code to be able know what to select as a marker and what not to.
- The method used to select the information to be extracted is unreliable and not exact. An alternative method for extracting information form HTML code from the source code of the web page has to be found in order to make the application more robust and useful.
- The user is not given a display that is comprised of all the information that they have extracted, instead the information is held in a number of different files. However using Cold Fusion to "CF include" can combine all scraped files into a template page. This is not practical; an alternative method of display has to be developed.
- Under the current implementation of the scraper all images that have a relative path are not displayed. A method of changing relative paths to absolute paths within the scraped information needs to be found.
- The granularity of the information that can be extracted reliably is quite high. If the user only wants to extract a small part of a URLs source HTML it becomes difficult if not impossible using this method of extraction. An alternative method would hopefully improve the reliability and range of granularity of information that can be extracted.
- There is no way for users to store information that they have scraped as data. Storage of data in a spreadsheet or database could be a valuable addition to this application.

3.3 Summary

Screen scraping is an excellent method of performing general information aggregation. This initial application extracts information in a crude manner and so is interesting as an example of one way that screen scraping or HTML extraction can be accomplished. A personal computer based application or web based application would have to perform the task of screen scraping in a much more elegant fashion than this one to be of any real use to the user. However the value of screen scraping to a WAP based web application provides motive to investigate alternative methods of information extraction.

4. DOM SCRAPER

Based upon the findings that the initial screen scraping application uncovered, alternative methods of information extraction were researched. The most

promising method for reliable information extraction to date is to use the Document Object Model (DOM, 3Com 1999) to navigate the HTML tree in order to locate and extract information that a user might want. There are a number of ways in which this can be achieved. One appealing method is to use the Microsoft Internet Explorer Control, which exposes the Document Object Model interface to a Visual Basic application.

The IE control can be integrated into a web based application. This results in an infinitely improved application in terms of flexibility and accessibility. The granularity of information extraction is improved dramatically by using the DOM, which gives you access to almost any element in a web page. Also, the speed in which information can be extracted using the DOC interface exposed by the IE control is dramatically faster than any other methods.

4.1 Process

The Internet Explorer control allows the user to click on some information that they would like to scrape. From that click you can find out which table holds the information that the user wants, change its border size and colour to indicate to the user information held in said table.

Figure 3 shows an example of an ActiveX control that is being used to aid users in the selection of information from a web page. The user types or pastes the URL that they want to extract information from into the ActiveX control, which essentially acts as a Web Browser. The main component of the ActiveX is the Web Browser Control, which enables

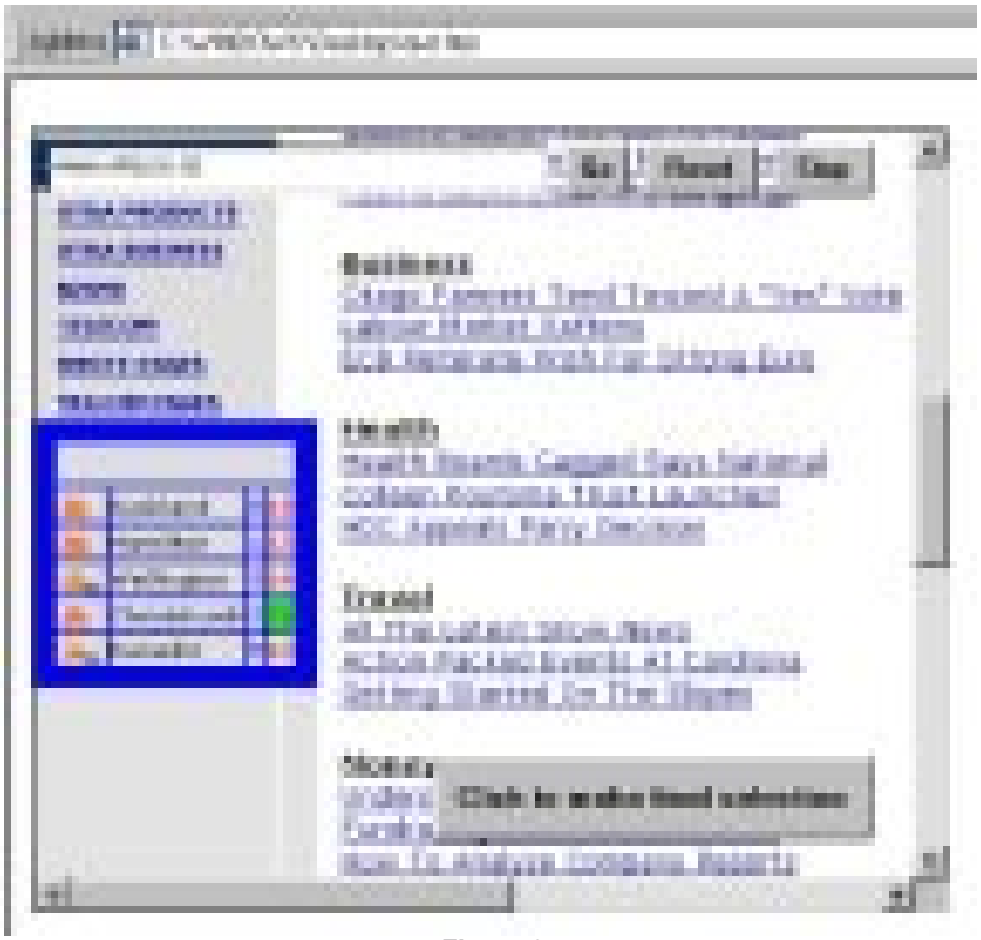

Figure 3

The ActiveX component shows the page to be scraped. The user has clicked on Dunedin, the table structure is shown and the chosen cell highlighted.

the Document Object Model to be used to extract information. This prototype allows information to be extracted from tables with in a web page. Most commercial dynamic web pages use tables to arrange information.

To select information from the web page the user clicks on it with the mouse. Using the DO, the outer HTML of the element is captured and it is located in the HTML of the web page. The table border and the background colour of the selected cell is changed providing the user with a visual representation of the information that has been selected. At this point the Final selection button appears in the ActiveX control, once clicked the information selected by the user is formatted and displayed in a similar format to that which will be displayed on the phone (Figure 4).

The user selects either the contents of the table or the contents of the selected table cell. The ActiveX control has now finished and methods expose data that it has collected to the outside world.

The following metadata is collected by the application:

- URL
- Table index number
- Row index number
- Cell index number

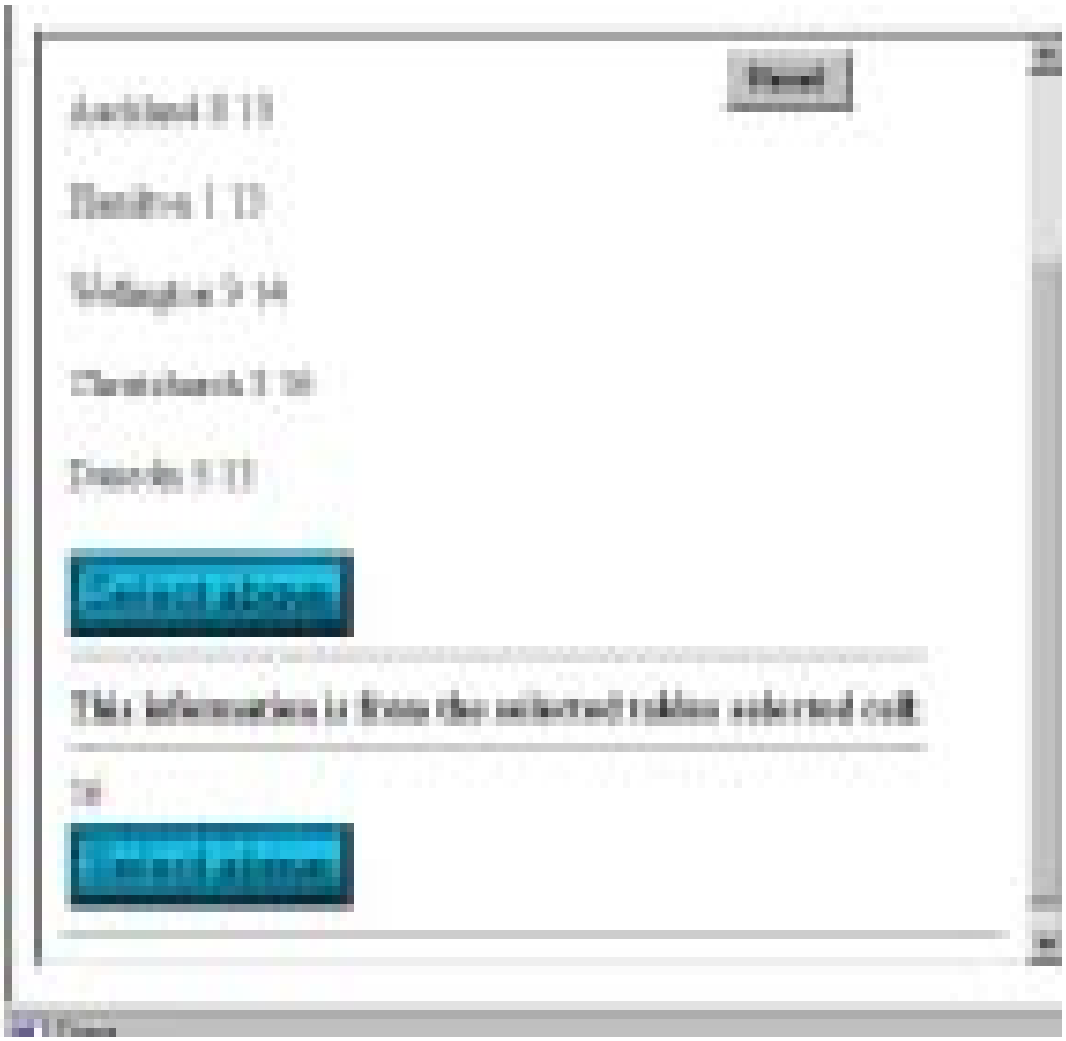

Figure 4

Using the DOM the selected piece of information is identified and shown to the user for final confirmation. It is the metadata that is stored, not the data value itself.

This information can be inserted into a database that stores the above data using XML. This is then available via Cold Fusion to a web page to be selected and ranked. When called by cell phone, a retrieval COM object returns and formats information into WML suitable for WAP phone. There are then, two sites, an HTML site used for administration and information gathering. A WML site that allows users to access gathered information.

Once a user has signed on as a member of the site the following services will be available to them.

- Ability to create groups each containing a number of scrapes, each scrape will retrieve information from targeted HTML pages and present them as cards on a WAP device.
- The ability to add, edit, delete, rename, groups or any given group members.
- The ability to directly access groups of groups, or individual groups via a WAP device displaying contained information on the WAP device.

5. DISCUSSION

The problem of information availability is exacerbated by user's increasing reliance on WAP/WML technology as a primary information device. This project has demonstrated it is possible to develop a web based application that allows users to scrape information from the web and access it from a WAP cell phone.

This process uses the DOM to identify metadata that describe the location of information. This metadata is described as XML and used to dynamically generate WML when requested by a user for display on a cell phone.

The following areas required further development:

- Ability of users to choose multiple sections of a web page as one target scrape
- Improve the formatting of targeted text when converting from HTML to WML
- Resolve problems that some frame sites cause
- Commercializing the interfaces and web applications beyond the prototype applications described here.

Vartanian and Ledig (2001) discuss in some detail the legal ramifications of data aggregation involving personal data such as financial records. In addition to the security and privacy concerns, other issues focus on the erosion of the brand (being scraped) and the loss of control over the user's web experience, which, for example, reduces opportunities to cross sell. These issues come under the area of copyright. Taking information from one website and reformulating it on another involves a significant amount of copying, certain aspects of which may be protected under the copyright laws. If the aggregator merely takes hard factual data and presents it on its own website in its own format there can be less objection on copyright grounds than if it adopts aspects of the source sites "look and feel". This requires further investigation as to how it applies to information delivered to WML for mobile use.

REFERENCES

3Com (1999) Document Object Model (DOM) Level 1 Specification http://www.w3.org/TR/REC-DOM-Level-1/ Accessed 18/5/01

Nielsen, J. (1999) Metcalfe's Law in Reverse http://www.useit.com/alertbox/990725.html Accessed 18/5/01

Nielsen, J. (2000) WAP Backlash. Nielsen's Alertbox http://www.useit.com/alertbox/20000709.html Accessed 8/5/01

Hackathorn, R. D. (1999) Webfarming Newsletter (online) Accessed 22/11/00 http://www.webfarming.com/new/NL199907a.html

Sahuguet, A.and Azavant, F. (1999) Wysiwyg Web Wrapper Factory (W4F) University of Pennsylvania http://db.cis.upenn.edu/Publications/ Accessed 21/11/00

Sorme, K. (2001) Internet in your pocket by Ericcson. Reported online Stuff 23/01/01 http://www.stuff.co.nz/index/0,1008,604065a1892,FF.html

Vartanian, T.P. and Ledig R.H. (2001) Scrape it, scrub it and show it: the battle over data aggregation. http://www.ffhsji.com/bancmail/bmarts/aba_art.htm Accessed 23/11/00

ACKNOWLEDGEMENT

This work was funded by an Otago Polytechnic School of Information Technology and Electrotechnology Research Grant.