Data Mining on the Web

Sharanabasavaraj B. Patil, Vijay Kulkarni, S. Chandrashekhar and M. H. Sunil

February 12, 2020

DATA MINING ON THE WEB

Sharanabasavaraj B. Patil[1], Vijay Kulkarni[2], Chandrashekhar S[3], Sunil M H[4]

1. Department of Physics, Government First Grade College, Ranebennur – 581115, Karnataka
2. Department of Engineering Physics Angadi Institute of Technology & Management, Belagavi, Karnataka
3. Department of Computer Science, Government First Grade College, Raichur, Karnataka
4. Department of Physics, Government First Grade College, Ranebennur – 581115, Karnataka

Corresponding Author Mail : sharanub.patil@gmail.com

Abstract : Data Mining [1, 2] or knowledge discovery is a method of extracting interesting, non-trivial, implicit, previously unknown and potentially useful information or patterns of data from large databases**.** Data mining is a computer-assisted process of digging and analyzing enormous sets of data and then extracting the desired information or data. Data mining tools also predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions.

There are two types of data mining techniques; ***descriptive data mining*** that describes the general properties of the existing data, and ***predictive data mining*** that attempts to do predictions based on inferences on available data. The patterns obtained are used to describe concepts, to analyze associations, to build classification and regression models, to cluster data, to model trends in time-series and to detect outliers ("data objects that do not comply with the general behavior or model of the data"). Since the patterns which are present in data are not all equally useful and interesting, measures are needed to estimate the relevance of the discovered patterns to guide the mining process. The data mining software performs an analysis of relationships and patterns in stored transaction data based on open-ended user queries.

Keywords: IE, MRD, DEPTA, VSAP, HTML, Bounded Rectangle, Data Record, Data Region,

Introduction :

Existing Techniques : Web information extraction (IE) from the data region is an important problem. Identifying the data region is often the first step. So far, several attempts have been made to deal with the problem. Earlier works on IE are mainly semi-automatic or even manual. They rely on training and human assistance to generate extraction rules for web pages. Several automated or semi-automated methods have been proposed recently and the most representative ones are Mining Data Records (MDR), Data Extraction using Partial Tree Alignment (DEPTA), Vision Based Page Segmentation (VIPS).

Visual Structure based Analysis for mining web Pages (Our Method of Data Extraction) :

Most of the existing techniques are unable to identify the relevant data without accessing the content of the web page. An important observation with regard to web pages is that area of the bounding rectangle of the data region is more than the area of other bounding rectangles. If an effective technique is developed for identifying the data region in a web page, data mining will become relatively easier and faster.

The system model of VSAP is shown in Fig-1. The components included in VSAP are **data region extractor** and **data extractor**. The data region extractor in turn has three components

1. Largest Rectangle Identifier
2. Container Identifier
3. Data Region Identifier

The output of each of these components is the input to the next component.

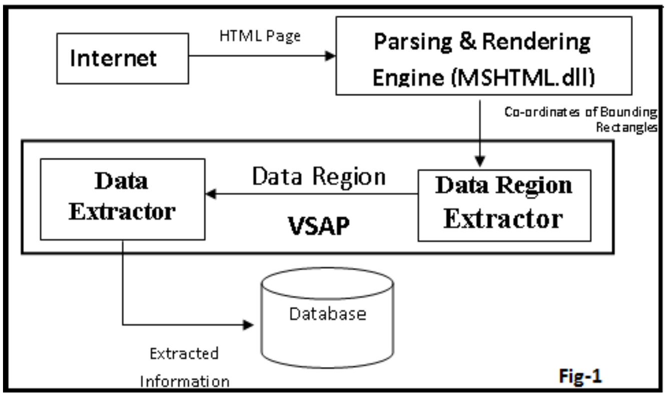


Fig-1

The VSAP method is based on three observations:

1. A group of data records that contains descriptions of a set of similar objects are typically presented in a contiguous region of a page. For example, in Fig-2 four books are presented in one contiguous region. The information concerning each book forms a data record.
2. The area covered by a rectangle that bounds the data region is more than the area covered by rectangles bounding other regions, e.g. advertisements and links.
3. The Height of an irrelevant data record within a collection of data records is less than the average height of relevant data records within that region.

Given a web page, the proposed technique works in three steps:
Step 1) Determining the co-ordinates of all bounding rectangles in the web page.
Step 2) Identify the data region.
Step 3) Extract data from the data region.

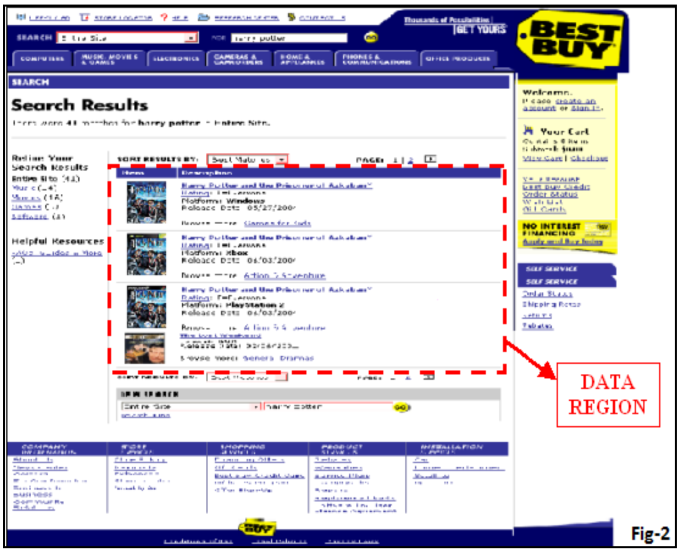


Fig-2

Determining Co-ordinates of all Bounding Rectangles

A Bounding Rectangle : Every HTML tag specifies a method for rendering the information contained within it. For each tag, there exists an associated rectangular area on the screen. Any information contained within this rectangular area obeys the rendering rules associated with the tag. This rectangle is called the bounding rectangle for the particular tag, Fig-3 shows bounding rectangles for all the TD tags in the webpage.

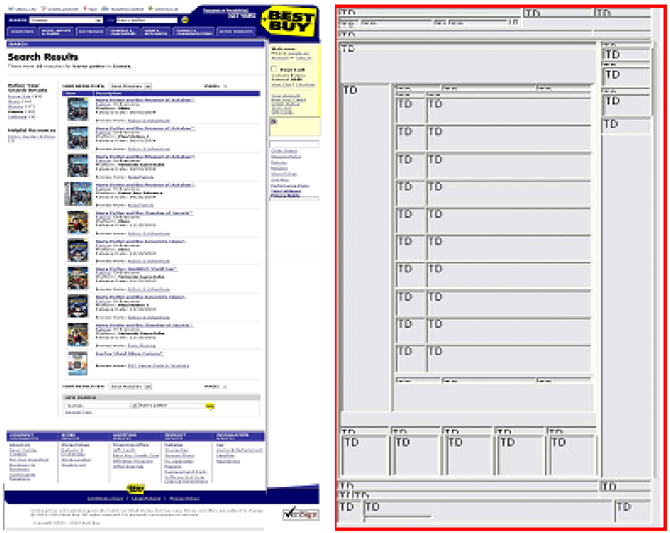


Fig-3

A bounding rectangle is constructed by obtaining the co-ordinate of the top-left corner of the tag, the height and the width of that tag. The left and top co-ordinates of the tag are obtained from the offset-left and offset-top properties of the HTML object element. These values are with respect to its parent tag. The height and width of that tag are available from the offsetHeight and offsetWidth properties of the HTMLObjectElement Class. Fig-4 shows the top-left corner co-ordinate, height and width obtained for constructing the bounding rectangle of an example tag. The bounding rectangle of that tag is shaded in red.

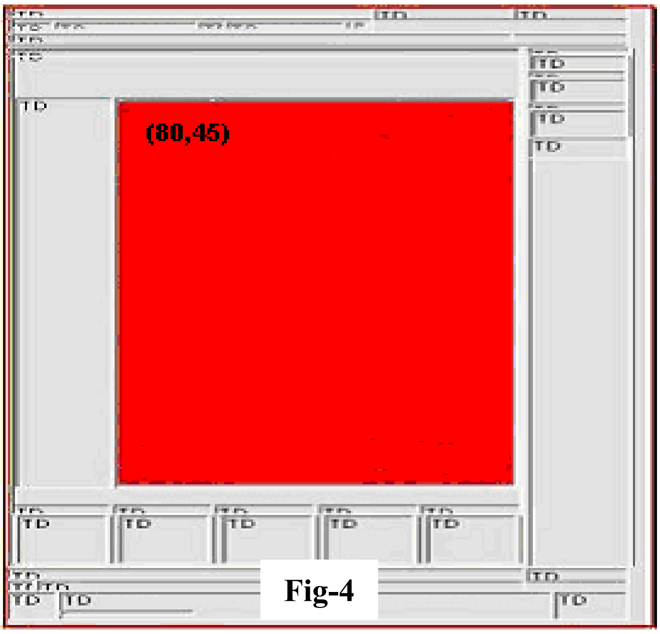


Fig-4

A web browser has a parsing and rendering engine, whose function is to parse the HTML file and render or display the information in a graphical way on the screen. The parsing and rendering engine of the web browser gives us these properties of a bounding rectangle. The VSAP approach uses the MSHTML parsing and rendering engine.
Whenever internet explorer (IE) encounters an HTML document, IE with the help of MSHTML parses then renders the HTML document appropriately on the browser window. We scan the HTML file for tags. For each tag encountered, we determine the co-ordinate of the top-left corner, height and width of the bounding rectangle of that tag.

Step -2) Identify the data region :

The data region is the most relevant portion of a web page that contains a list of contiguous records. There are 3 steps involved in identifying the data region, which are :

a) Identify the largest rectangle.
b) Identify the container within the largest rectangle.
c) Identify the data region within this container

Identifying the largest rectangle

Based on the height and width of bounding rectangles obtained in the previous step, we determine the area of the bounding rectangles of each of the children of the BODY tag. We then determine the largest rectangle amongst these bounding rectangles. The reason for doing this is due to the observation that the largest bounding rectangle will always contain the most relevant data in that web page. Thus by determining the largest rectangle, we can thereby obtain a superset of the data region.

Identify the container within the largest rectangle

Once we have obtained the largest rectangle, we form a set of all the bounding rectangles whose area is more than half the area of the largest rectangle. The rationale behind this is that the most important data of a web page must occupy a significant portion of the web page. We next determine the bounding rectangle having the smallest area in this set. The reason for determining the smallest rectangle within this set is that the smallest rectangle will only contain data records. Thus a container is obtained, which contains the data region and some irrelevant data. A *container* can be defined as superset of the data region which may or may not contain irrelevant data. Fig-5 shows the container identified from the web page given. Note that the advertisements on the right and bottom of the page and the links on the left side are removed.

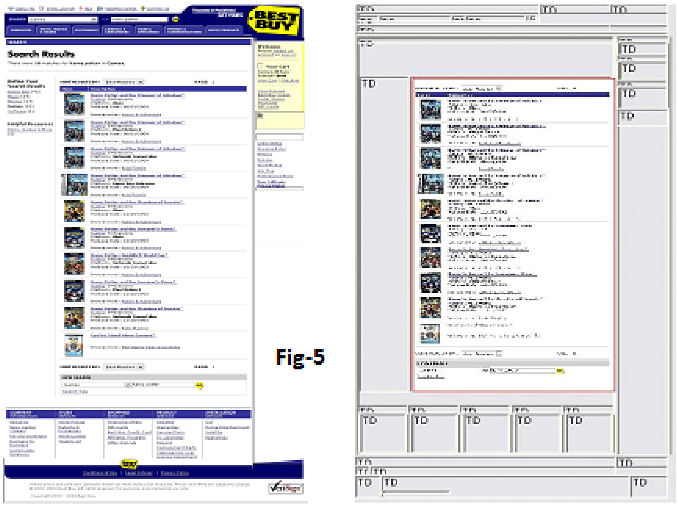


Fig-5

Identify the data region within this container

Now to filter the irrelevant data from the container, we use a filter. The filter determines the average heights of children within the container. Those children whose heights are less than the average height are identified as irrelevant data and are filtered off. Fig-6 shows a filter applied on the container to the left, to obtain the data region on the right. Note that the irrelevant data in this case is on the top and bottom of the filter.

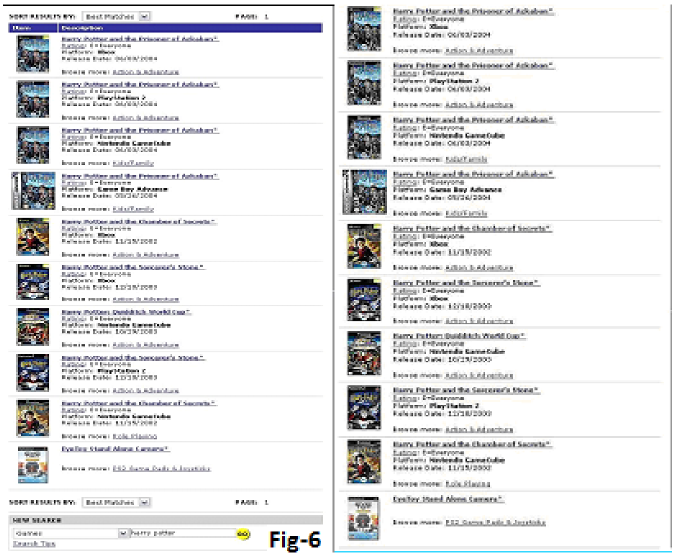


Fig-6

Step 3 : Extract data from the Data Region :

The final step in VSAP is to extract data from data records from the data region. The data from the data fields from each data record is stored into a database. Thus the output of this step is a table consisting of data fields and the data for each of these fields.

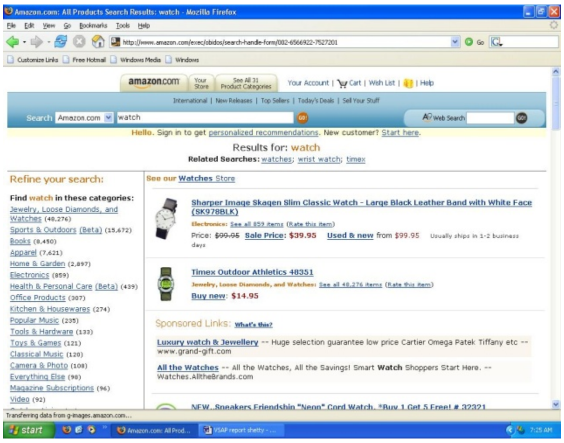
If VSAP is applied on a number of web-pages, data from the data region of each of these web-pages can be extracted. Data extracted from each web-page can then be integrated into a single collection. This collection of data can be further used for various purposes like making a comparative study of products from various companies, smart shopping, etc. Thus the data collected can be used for knowledge discovery applications.

Conclusion :

A novel and effective method to mine the data region in a web page is developed. The method is based on recognizing data regions using HTML tags. This method called as visual structure based analysis of web pages (VSAP) is a pure visual structure oriented method that can correctly identify the data region. VSAP is independent of errors occurring due to misuse of HTML tags. Also most current algorithms fail to correctly determine the data region, when the data region consists of only one data record. Also, most sites fail in the case where a series of data records is separated by an advertisement, followed again by a single data record. It works correctly for both the above cases. It is able to correctly identify data records, irrespective of the type of tag in which it is bound. Other methods would work for only tags of certain types. The number of comparisons done in VSAP is significantly lesser than other approaches. Also comparisons are made on numbers, unlike other methods where strings or trees are compared. VSAP overcomes the drawbacks of most existing methods and has correctly identified the data region on all pages tested.

Comparing our method with MDR method

Amazon.com



Amazon.com: All Products Search Results: watch - Mozilla Firefox

http://www.amazon.com/exec/obidos/search-handle-form/002-6566922-7527201

Your Account | Cart | Wish List | Help

International | New Releases | Top Sellers | Today's Deals | Sell Your Stuff

Search Amazon.com watch

Hello. Sign in to get personalized recommendations. New customer? Start here.

Results for: watch

Related Searches: watches; wrist watch; timex

Refine your search:

Find watch in these categories:

Jewelry, Loose Diamonds, and Watches	(48,276)
Sports & Outdoors (Beta)	(15,672)
Books	(8,450)
Apparel	(7,621)
Home & Garden	(2,897)
Electronics	(859)
Health & Personal Care (Beta)	(439)
Office Products	(307)
Kitchen & Housewares	(274)
Popular Music	(235)
Tools & Hardware	(133)
Toys & Games	(121)
Classical Music	(120)
Camera & Photo	(108)
Everything Else	(98)
Magazine Subscriptions	(96)
Video	(92)

See our Watches Store

Sharper Image Skagen Slim Classic Watch - Large Black Leather Band with White Face (SK978BLK)

Electronics: See all 859 items (Rate this item)

Price: ~~$99.95~~ Sale Price: $39.95 Used & new from $99.95 Usually ships in 1-2 business days

Timex Outdoor Athletics 48351

Jewelry, Loose Diamonds, and Watches: See all 48,276 items (Rate this item)

Buy new: $14.95

Sponsored Links: What's this?

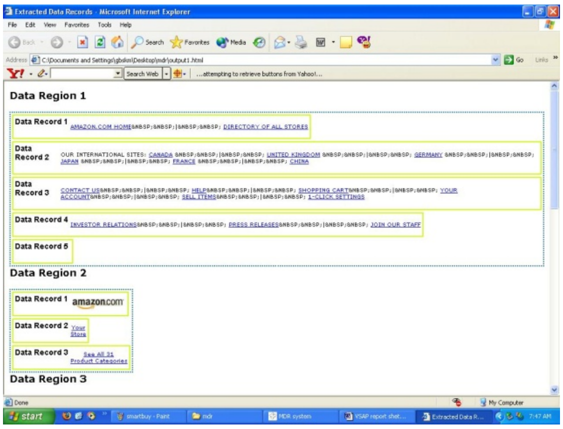
Luxury watch & Jewellery -- Huge selection guarantee low price Cartier Omega Patek Tiffany etc -- www.grand-gift.com

All the Watches -- All the Watches, All the Savings! Smart Watch Shoppers Start Here. -- Watches.AlltheBrands.com

NEW..Sneakers Friendship "Neon" Cord Watch. *Buy 1 Get 5 Free! # 32321

Transferring data from g-images.amazon.com...

Result of MDR



Extracted Data Records - Microsoft Internet Explorer

C:\Documents and Settings\gbskm\Desktop\mdr\output1.html

Data Region 1

Data Record 1 AMAZON.COM HOME&NBSP;&NBSP;|&NBSP;&NBSP; DIRECTORY OF ALL STORES

Data Record 2 OUR INTERNATIONAL SITES: CANADA &NBSP;&NBSP;|&NBSP;&NBSP; UNITED KINGDOM &NBSP;&NBSP;|&NBSP;&NBSP; GERMANY &NBSP;&NBSP;|&NBSP;&NBSP; JAPAN &NBSP;&NBSP;|&NBSP;&NBSP; FRANCE &NBSP;&NBSP;|&NBSP;&NBSP; CHINA

Data Record 3 CONTACT US&NBSP;&NBSP;|&NBSP;&NBSP; HELP&NBSP;&NBSP;|&NBSP;&NBSP; SHOPPING CART&NBSP;&NBSP;|&NBSP;&NBSP; YOUR ACCOUNT&NBSP;&NBSP;|&NBSP;&NBSP; SELL ITEMS&NBSP;&NBSP;|&NBSP;&NBSP; 1-CLICK SETTINGS

Data Record 4 INVESTOR RELATIONS&NBSP;&NBSP;|&NBSP;&NBSP; PRESS RELEASES&NBSP;&NBSP;|&NBSP;&NBSP; JOIN OUR STAFF

Data Record 5

Data Region 2

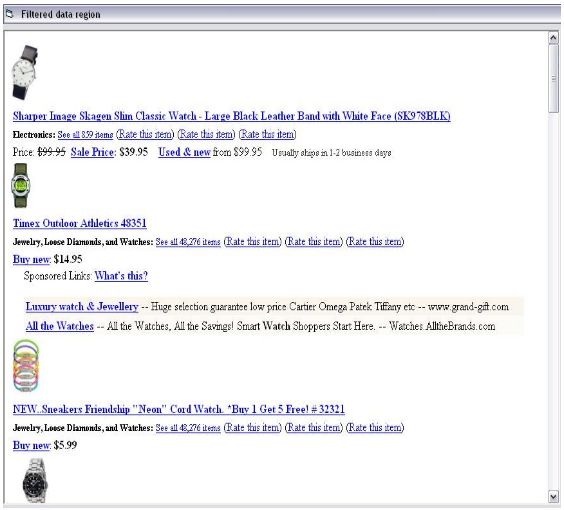
Data Record 1 amazon.com

Data Record 2 Your Store

Data Record 3 See All 31 Product Categories

Data Region 3

Result of VSAP



Filtered data region

Sharper Image Skagen Slim Classic Watch - Large Black Leather Band with White Face (SK978BLK)

Electronics: See all 859 items (Rate this item) (Rate this item) (Rate this item)

Price: ~~$99.95~~ Sale Price: $39.95 Used & new from $99.95 Usually ships in 1-2 business days

Timex Outdoor Athletics 48351

Jewelry, Loose Diamonds, and Watches: See all 48,276 items (Rate this item) (Rate this item) (Rate this item)

Buy new: $14.95

Sponsored Links: What's this?

Luxury watch & Jewellery -- Huge selection guarantee low price Cartier Omega Patek Tiffany etc -- www.grand-gift.com

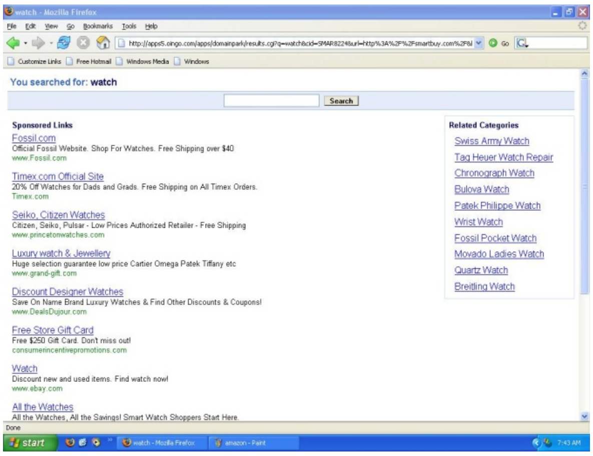
All the Watches -- All the Watches, All the Savings! Smart Watch Shoppers Start Here. -- Watches.AlltheBrands.com

NEW..Sneakers Friendship "Neon" Cord Watch. *Buy 1 Get 5 Free! # 32321

Jewelry, Loose Diamonds, and Watches: See all 48,276 items (Rate this item) (Rate this item) (Rate this item)

Buy new: $5.99

Smartbuy.com



watch - Mozilla Firefox

http://apps5.oingo.com/apps/domainpark/results.cgi?q=watch&cid=SMAR8224&url=http%3A%2F%2Fsmartbuy.com%2F&l

You searched for: watch

Search

Sponsored Links

Fossil.com
Official Fossil Website. Shop For Watches. Free Shipping over $40
www.Fossil.com

Timex.com Official Site
20% Off Watches for Dads and Grads. Free Shipping on All Timex Orders.
Timex.com

Seiko, Citizen Watches
Citizen, Seiko, Pulsar - Low Prices Authorized Retailer - Free Shipping
www.princetonwatches.com

Luxury watch & Jewellery
Huge selection guarantee low price Cartier Omega Patek Tiffany etc
www.grand-gift.com

Discount Designer Watches
Save On Name Brand Luxury Watches & Find Other Discounts & Coupons!
www.DealsDujour.com

Free Store Gift Card
Free $250 Gift Card. Don't miss out!
consumerincentivepromotions.com

Watch
Discount new and used items. Find watch now!
www.ebay.com

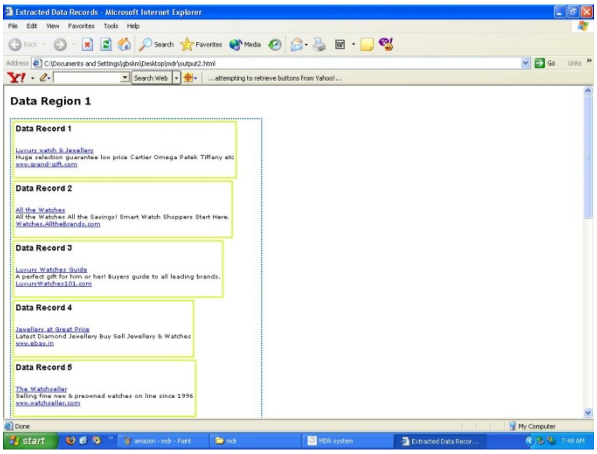
All the Watches
All the Watches, All the Savings! Smart Watch Shoppers Start Here.

Related Categories

- Swiss Army Watch
- Tag Heuer Watch Repair
- Chronograph Watch
- Bulova Watch
- Patek Philippe Watch
- Wrist Watch
- Fossil Pocket Watch
- Movado Ladies Watch
- Quartz Watch
- Breitling Watch

Done

Result of MDR



Extracted Data Records - Microsoft Internet Explorer

C:\Documents and Settings\gbskm\Desktop\mdr\output2.html

Data Region 1

Data Record 1
Luxury watch & Jewellery
Huge selection guarantee low price Cartier Omega Patek Tiffany etc
www.grand-gift.com

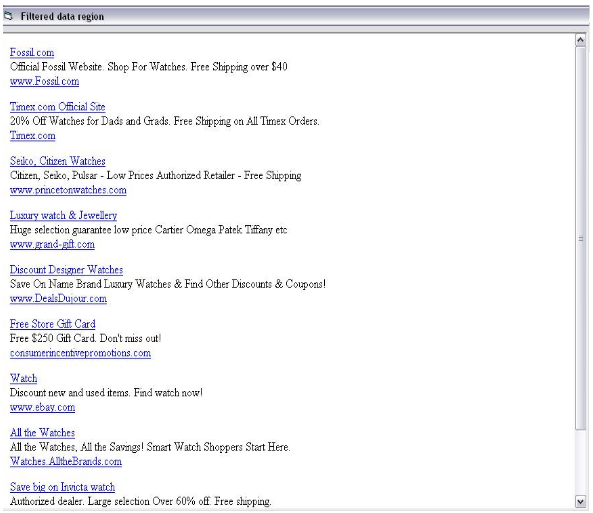
Data Record 2
All the Watches
All the Watches All the Savings! Smart Watch Shoppers Start Here.
Watches.AlltheBrands.com

Data Record 3
Luxury Watches Guide
A perfect gift for him or her! Buyers guide to all leading brands.
LuxuryWatches101.com

Data Record 4
Jewellery at Great Price
Latest Diamond Jewellery Buy Sell Jewellery & Watches
www.ebay.in

Data Record 5
The Watchseller
Selling fine new & preowned watches on line since 1996
www.watchseller.com

Result of VASP



Filtered data region

Fossil.com
Official Fossil Website. Shop For Watches. Free Shipping over $40
www.Fossil.com

Timex.com Official Site
20% Off Watches for Dads and Grads. Free Shipping on All Timex Orders.
Timex.com

Seiko, Citizen Watches
Citizen, Seiko, Pulsar - Low Prices Authorized Retailer - Free Shipping
www.princetonwatches.com

Luxury watch & Jewellery
Huge selection guarantee low price Cartier Omega Patek Tiffany etc
www.grand-gift.com

Discount Designer Watches
Save On Name Brand Luxury Watches & Find Other Discounts & Coupons!
www.DealsDujour.com

Free Store Gift Card
Free $250 Gift Card. Don't miss out!
consumerincentivepromotions.com

Watch
Discount new and used items. Find watch now!
www.ebay.com

All the Watches
All the Watches, All the Savings! Smart Watch Shoppers Start Here.
Watches.AlltheBrands.com

Save big on Invicta watch
Authorized dealer. Large selection Over 60% off. Free shipping.

Advantages of VSAP

a) *Pure Visual Structure Oriented Method*.
VSAP is purely dependent on the visual structure of the web page only. It does not make use of the text content of the web page at any stage. This is advantageous as it saves us from the additional overhead of performing keyword search on the web page.

b) *The entire tag-tree need not be scanned*.
Only the largest child of the BODY tag is scanned. The remaining part of the web-page need not be accessed. This is a very efficient method with complexity much lesser than other contemporary algorithms.

c) *VSAP is independent of errors occurring due to misuse of HTML tags*.
An incorrect tag tree may be constructed due to misuse of HTML tags in MDR, however there is no possibility of this happening in VSAP. This is because, in VSAP, the hierarchy of tags is constructed based on the visual cues on the web page.

d) *Data Region identification is independent of specific tags*.
A data record or data region need not be contained is specific tags like TABLE, TBODY, etc. Certain methods like MDR and DEPTA are dependent on certain tags for identifying data records.

e) *Comparison is made on numbers*.
In MDR, comparison of generalized nodes is made on strings. In DEPTA, comparison of sub-trees is made by tree matching. However, both these methods are slow and inefficient as compared to the number comparison made in VSAP. Number comparison is made in VSAP when comparing the co-ordinates of two bounding rectangles.

f) *Complexity of VSAP much lesser than existing algorithms*.
Other existing algorithms have a complexity of the order of greater than m*n while VSAP has a complexity of the order of n, where n is the number of tag-comparisons made and m is the depth of the tag tree.

REFERENCES

1. Bing Liu, Robert Grossman, Yanhong Zhai. '*Mining Data Records in Web Pages'*, ACM[1] SIGKDD[2] International Conference on Knowledge Discovery & Data Mining (KDD[3]- 2003), Washington, DC, USA, August 24 - 27, 2003. http://www.cs.uic.edu/~liub, http://citeseer.ist.psu.edu/article/liu03mining.html)
2. Soumen Chakraborthy, '*Data Mining on the Web',* Published by Morgan Kauffman, 1st Edition, August 15th 2002.
3. Deng Cai, Shipeng Yu, Ji-Rong Wen and Wei-Ying Ma Extracting content structure for web pages based on visual representation, Proc. 5th Asia Pacific Web Conference, Xi'an China, 2003. (URL : www.dbs.informatik.uni-muenchen.de)
4. Bing Liu, Kevin Chen-Chuan-Chang, Editorial : special issue on web content mining, ACM SIGKDD Explorations Newsletter, v.6 n.2, p.1-4, December 2004.
5. A. Arasu, H. Garcia-Molina, Extracting structured data from web pages, ACM SIGMOD[4] 2003 (URL : http://dblp.uni-trier.de/db/conf/sigmod/sigmod2003.html#ArasuG03)
6. Hongkun Zhao , Weiyi Meng , Zonghuan Wu , Vijay Raghavan , Clement Yu, Fully automatic wrapper generation for search engines, Proceedings of the 14th international conference on World Wide Web, May 10-14, 2005, Chiba, Japan.
7. Valter Crescenzi, Giansalvatore Mecca, Paolo Merialdo, RoadRunner : Towards Automatic Data Extraction from Large Web Sites, Proceedings of the 27th International Conference on Very Large Data Bases, p. 109-118, September 11-14, 2001.
8. Yanhong Zhai , Bing Liu, Web data extraction based on partial tree alignment, Proceedings of the 14th international conference on World Wide Web, May 10-14, 2005, Chiba, Japan
9. Baeza Yates, R. Algorithms for string matching : A survey. ACM SIGIR[5] Forum, 23(3- 4):34-58, 1989. (URL : http://www.acm.org/sigs/sigir/forum/index.html)
10. Yudong Yang, Hong Jiang Zhang, "HTML Page Analysis Based on Visual Cues" p.0859, Sixth International Conference on Document Analysis and Recognition (ICDAR) 2001.
11. D. Buttler, L. Liu, C. Pu. A Fully Automated Object Extraction System for the World Wide Web. International Conference on Distributed Computing Systems (ICDCS) 2001.
12. Chia-Hui Chang , Shao-Chen Lui, IEPAD: Information Extraction based on PAttern Discovery, Proceedings of the 10th international conference on World Wide Web, p.681-688, May 01-05, 2001, Hong Kong.

[1] Association for Computing Machinery.

[2] Special Interest Group Knowledge Discovery and Data Mining.

[3] Knowledge Discovery and Data Mining.

[4] Special Interest Group on Management of Data.

[5] Special Interest Group on Information Retrieval.