Extracting Semantic Structure of Web Pages Using Graph Grammar Induction Algorithm

B.Venkatesh, P.Prakash

Abstract-With the appearance of the web, it's fascinating to interpret and extract helpful data from the net. One major challenge in internet interface interpretation is to get the semantic structure underlying an internet interface. Several heuristic approaches have been developed to get and cluster semantically related interface objects. However, those approaches cannot solve the problem of non similarity satisfactorily and don't seem to be ready to tag the participant role of every object. Distinct from existing approaches, this paper develops a sturdy and formal approach to ill interface semantics mistreatment graph grammars induction. Due to the distinct capability of spatial specifications within the abstract syntax, the spatial graph grammar induction algorithm (SGGI) is chosen to perform the semantic grouping and interpretation of divided screen objects. Instead of analyzing HTML supply codes, we tend to apply an economical image processing technology to acknowledge atomic interface objects from the screenshot of an interface and manufacture a spatial graph, which records vital spatial relations among recognized objects. A spatial graph is a lot of taciturn than its corresponding document object model structure and, thus, facilitates interface analysis and interpretation. Supported the spatial graph, the SGGI parser recovers the graded relations among interface objects.

Keywords - Content extraction, Image Segmentation, Graph Grammar Induction Algorithm, Spatial Parsing.

I INTRODUCTION

With the big quantity of heterogeneous knowledge on the Web, it is fascinating to mechanically interpret a Web interface and extract helpful data. One major challenge in Web interface interpretation is to find the net interface linguistics, i.e., page segmentation, that teams semantically connected interface objects during a hierarchical data structure and consequently tags the syntactic category of every object. Since net interfaces square measure created autonomously, the irregularities caused by completely different designers and organizations build it difficult to extract interface linguistics. Several researchers have explored heuristic approaches [1], [7], [10], to discovering (the data the knowledge the data) organization underlying a Web page.

These heuristic approaches in general perform page segmentation by analyzing the document object model (DOM) structure through a collection of heuristic rules. However, markup language could be a terribly versatile language and completely different designers might use the markup language fully otherwise. For example, tables in markup language square measure designed to prepare and show tabular knowledge, implying that data during a table is closely related. However, by not displaying the table border, several developers use a table as a company grid to layout photos and texts. during this case, data fencelike during a table might not essentially be semantically relevant. Ahmadi and Kong [1] have evaluated six heuristic rules on three genres of websites (e.g., news, travel, and shopping) and ended that heuristic rules have accuracies on different genres. Additionally to the variety of markup language usages, the quality of DOM structures conjointly negatively affects the performance of page segmentation. Furthermore, heuristic approaches will cluster closely related data. However they're ineffectual of tagging linguistics roles.

Recently, visual language formalisms are projected to analyze Web interfaces. Distinct from heuristic approaches, this approach formalizes a standard net pattern as a graph grammar, that formally and visually specifies the data organization underlying an internet page. The grammar-based approach interprets an internet page from bottom to high and, thus, needs to 1st acknowledge atomic data objects before page segmentation. However, it's difficult to acknowledge atomic data objects from the DOM structure. For instance, in an HTML Web page, a line of texts (i.e., associate degree atomic interface object) may be separated by many markup language tags, and therefore the seperation is content dependent.

Based on the preliminary work [12], this paper proposes a novel approach to page segmentation, taking advantage of graph grammars to supply strong page segmentation while not relying on DOM structures. thanks to the distinctive spatial specification capability within the abstract syntax, the spatial graph synchronic linguistics (SGG) [10] is employed in our approach to investigate net interfaces. Spatial specifications within the abstract syntax modify designers to model interface linguistics with varied visual effects (e.g., a topological relation between 2 interface objects). Our approach interprets an internet page, or any interface page, directly from its image, rather than DOM structures. Image-processing techniques [16] square measure accustomed divide associate degree interface image into completely different regions and acknowledge and classify atomic interface objects, such as texts, buttons, etc., in every region. The article recognition produces a spatial graph, during which nodes represent recognized atomic objects and edges indicate some vital spatial relationships (such as bit and containment). Finally, the SGG parser parses the spatial graph to find the stratified relations among those interface objects supported a predefined graph grammar.

To our data, the aforesaid approach is that the 1st to combine image process with graph synchronic linguistics, that effectively addresses the problem of dissimilarity and simplifies page segmentation. The discovered interface linguistics is helpful in many Web-based applications, like content adaptation, information

Manuscript Received March, 2014.

B.Venkatesh, Department of Computer Science and Engineering, K.S.R.College of Engineering, Thiruchengode, India.

Ass Prof P.Prakash, Department of Computer Science and Engineering, K.S.R.College of Engineering, Thiruchengode, India.

retrieval, or usability analysis. for instance, based on the discovered linguistics, we will adapt the data presentation to mobile devices or retrieve records by combining the interface linguistics with an information model and a question language. Furthermore, the interface interpretation will verify whether or not completely different Web pages adapt to a standard pattern or not. Additionally to net interfaces, our work will be applicable to traditional GUI applications and is helpful to enrich public application programming interfaces to extract helpful knowledge.

II RELATED WORKS

As a difficult issue in internet interface interpretation, page segmentation has attracted a lot of attention. Distinct from previous work, this paper develops a completely unique approach, that integrates the advantages of image-processing and graph-grammar techniques. The image-processing technique with efficiency acknowledges information objects and abstracts the first web content as a epigrammatic abstraction graph, whereas graph synchronic linguistics provides a solid foundation for decoding an internet interface in terms of its abstraction configuration. Currently, numerous page segmentation methods are projected for various applications. Those methods are evaluated on different Web sites. So as to have Associate in nursing objective and quantitative comparison among completely different approaches, it's necessary to line up many benchmark internet sites for the analysis and comparison purpose. Especially, those benchmark Web sites ought to cowl completely different classes thus that we are able to compare each the accuracy and generality.

In order to permit users to seek out data of interest quickly, a good internet styleer usually observes design tips to render information on the net. For instance, some HTML tags are normally wont to imply a boundary between interface objects. Those observations encourage numerous heuristic approaches, which discover blocks of closely connected contents by analyzing the visual look or the HTML DOM structure of an internet page. Those approaches area unit automatic and economical in grouping relevant data. Completely different from our grammar-based approach, they lack a proper basis and don't recover the linguistics role of every interface object, that is beneficial in several applications.

Many heuristic approaches [1], [11] use HTML structural tags (like Table) to partition an internet page. Kaasinen et al. [28] projected Associate in Nursing HTML/WML conversion proxy server, which converts HTML-based internet contents to WML by mapping HTML structures to WML specifications. For instance, it converts Associate in Nursing HTML table to a WML table, Associate in nursing indexed sub tree or an inventory in step with the table size and viewing capability. Buyukkokten et al. [11], [12] divided an internet page into many semantic matter units through HTML tags, e.g., the tag P may function the boundary between 2 linguistics matter units. This methodology, however, solely focuses on texts while not supporting graphics. Smart View [11] used a fingernail to produce a visual summary of a page and partition a page into logic units according to table tags. Opera [13] offers a small-screen rendering technology, that stacks internet contents vertically to avoid horizontal scrolling. This methodology might incorrectly separate closely related contents and mix unrelated data along. The DOM-structure-based analysis is restricted by the quality of DOM structures.

Recently, visual analysis has attracted additional and additional attention. Yang and Zhang [14] evaluated the visual similarities of HTML contents, detected the pattern of visual similarity, and then generated a hierarchic illustration of the HTML page. Chen et al. [15] initial divided an internet page into many high-level data blocks in step with their sizes and locations, and, then, known express and implicit separators within each high level block. Supported the partition, a Web page is adapted to many subpages with a two-level hierarchy: a thumbnail at the highest level for Associate in Nursing index of contents and a group of subpages at the lowest level for elaborated reading. CMo [8] utilizes geometrical alignment of frames to section an internet page. Paterno and Zichittella [16] dynamically split the presentation of a desktop page by shrewd the price (e.g., the number of pixels of pictures or font sizes) of knowledge objects. The vision-based page segmentation (VIPS) [17], utilizes helpful visual cues and DOM structures to get the partition of a Web page at the linguistics level. Xia et al. [22] adjusted the VIPS algorithmic program to supply Associate in Nursing SP-tree that represents the hierarchical data structure of knowledge blocks during a web content. Hattori et al. [23] calculated the strength of connections between content components supported the structural depth of HTML tags and analyzed the layout to section a page. Ahmadi and Kong [1] analyzed each the DOM structure and therefore the visual layout to divide the first web content into many subpages, each including closely connected contents and appropriate for small-screen display. This approach supports automatic generation of a table of contents to facilitate the navigation between completely different subpages.

Visual language formalisms are applied to analyzing patterns of internet queries. Given a synchronic linguistics within the type of a variant of the attributed multi set synchronic linguistics, that specifies commonly used internet question patterns, a best-effort program analyzes a question kind by parsing the spacing of visual objects within the shape. This paper emphasizes on question interface, rather than a full web content. Kong et al. [18] uses SGG to research the linguistics of an internet page. This approach is based on DOM specifications, rather than a picture analysis.

HTML scraping [19] has been wide wont to scrape HTML Web pages. Supported predefined regular expressions, the HTML scraping technique will with efficiency extract helpful information from internet pages. Wrapper induction [1],is employed to extract structured data from Web pages or semi structured documents. Labsk´y et al. [20] combined the hidden mathematician models with image classification to extract structured data. Wong and Lam [21] proposed a completely unique framework, which might mechanically adapt a previously learned wrapper from a supply site to a brand new unseen web site within the same domain. Rather than analyzing the organization of all data during a web content, those approaches emphasize on extracting structured information in response to a submitted question. Some researchers [3], [4] analyzed the elaborated contents in Associate in Nursing HTML Web page to extract semi structured data. Ashraf and Alhajj [3] projected a completely unique approach, called ClusTex, that applies a bunch technique for data extraction from HTML

pages. This approach initial uses a bunch technique to divide data into clusters, that area unit then refined to eliminate digressive data. Ashraf et al. [4] later applied ClusTex to variety of websites from completely different domains. The analysis results show smart performance. Different from ClusTex, our approach interprets an internet page from its layout, rather than elaborated contents.

Recently, the perception technique has been applied to extract structured information, since it's freelance from the elaborated implementation underlying a Web page. These approaches [1], [12] primarily calculate the visual similarity among completely different Web pages to cluster semantically connected data. Zheng et al. [12] introduced a model freelance system to spot news articles supported visual consistency. This approach summarizes a group of visual options to gift news stories and then mechanically generates a template-independent wrapper based on those visual options. Chen and Xiao [24] projected a system to extract news stories supported perception. First, it identifies the areas that contain news stories supported content practicality, house continuity, and information continuity. After detective work the news areas, news stories area unit extracted based on the position, format, and linguistics. The two aforesaid approaches [6], [12] area unit restricted to extract news stories and aren't applicable to different domains. Distinct from those two approaches, our approach is general to completely different application domains.

III SYSTEM DESIGN

Fig.1 a pair of shows an outline of our approach. Page segmentation proceeds in 2 steps: 1) acknowledge interface objects in associate interface image; and 2) interpret the interface through graph grammars.

Our approach interprets associate interface from bottom to high. Therefore, the primary step is to acknowledge and classify interface objects. Hypertext mark-up language tags could separate associate atomic interface object into many items. For instance, so as to spotlight several words, a sentence, i.e., associate atomic interface object, may be separated into many items by the hypertext mark-up language tag. Since the separation is totally content dependent, it is difficult to derive some general rules to consolidate those items into atomic interface objects from the attitude of DOM structures. In our approach, image segmentation techniques, like line detection, text detection, and guide matching, are applied to spot interface objects in associate interface image. Additionally, seeing techniques are used to recognize the interface object sorts, like buttons, text boxes, images, etc. seeing and classification divide an interface into many regions associated have the subsequent advantages.

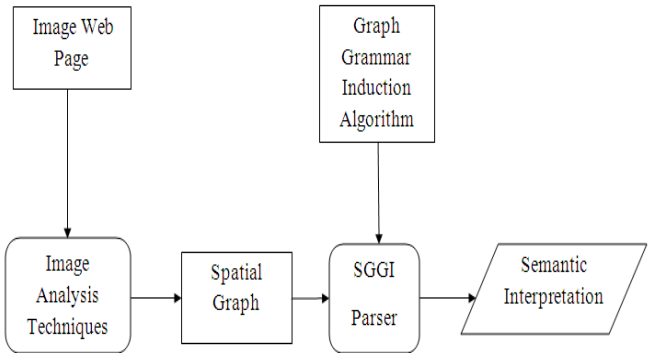

Fig.1System Architecture

A. progressive analysis

Those regions offer natural boundaries between composite interface objects. In other words, we are able to interpret interface objects incrementally first, interpret interface objects within one region, and then consider the relations between completely different regions. Such an incremental analysis is inherently in keeping with the hierarchical design of net interfaces. Designers, in general, first divide associate interface into many layout regions, and then confirm the organization and layout of interface objects at intervals every region.

B. Handling irregularity

If a vicinity contains vogue exceptions, not captured by a graph synchronic linguistics, we are able to merely interpret those exceptional interface objects because the direct children of the node representing the corresponding region.

C. Performance

Limiting the spatial parsing during a little region can cut back the search area and improve performance.

The visual analysis on the interface image generates a spatial graph. During a spatial graph, nodes represent recognized interface objects, and edges indicate significant spatial relations, each indicating an in depth linguistics relationship. We've got evaluated different Web sites and discovered that four spatial relations strongly imply an in depth linguistic relation between two objects, i.e., touching, containment, vertical, and horizontal relations within atiny low distance. Consequently, every come on a spatial graph corresponds to one of these four relations. Meanwhile, a spatial graph conjointly records the coordinates of every recognized interface object. Therefore, different further spatial relations will be derived from those coordinates. Compared with the supply HTML file, the spatial graph facilitates page segmentation by:1) consolidating info items together; and 2) removing all redundant contents (such as empty columns or tables, which are used for adjusting layout).

After visual analysis, a graph synchronic linguistics is applied to a spatial graph to find the interface linguistics. In our approach, page segmentation will be thought of as a graph transformation issue. The input could be a spatial graph that's abstracted from a concrete net interface, and therefore the output could be a tree that reveals the hierarchical relations among interface objects. Therefore, graph grammars are a natural process model for page segmentation. additional specifically, the left graph during a production may embody a composite interface object, that is formed of a collection of atomic/composite interface objects within the right graph. Each production teams connected objects regionally and a whole graph grammar provides a scientific specification to connect low-level groups into the next level cluster. Supported the graph synchronic linguistics, a computer program constructs a hierarchical parsing tree, during which a leaf node indicates associate atomic interface object associated an intermediate node represents a composite object, bottom up as a coherent interpretation for an internet interface. Our approach isn't restricted to a particular graph-grammar formalism. This paper uses the SGG [10] because the specification formalism to investigate net interfaces because of its distinct capability of spatial specification in the abstract

syntax. This distinctive feature permits USA to increase our approach to mix DOM analysis with image process in the future. The progressive parsing in our approach supports reusing some of a graph synchronic linguistics in numerous internet sites. Even though 2 websites from completely different internet sites is also different at a high level, some low-level patterns, like the organization of a paragraph, is also used repetitively across different internet sites.

Following the human–computer interaction principle that consistent layouts will improve the usability of associate interface, Web designers unremarkably use similar layouts to gift the same style of info. In different words, designers, in general, follow previous triple-crown experiences, which might be summarized as tips, to gift interface objects. Some researchers summarized common style patterns across completely different Web sites. We've got evaluated twenty one industrial internet sites and found that each one those internet sites use 2 common patterns to demonstrate product info. The usage of common patterns or customary tips makes our approach applicable in follow. A graph synchronic linguistics will be applied to completely different net sites that adjust to one common pattern or customary. In order to reduce the manual efforts of coming up with a graph synchronic linguistics, we attempt to introduce synchronic linguistics induction technique to automatic the synchronic linguistics style method within the future. the automated synchronic linguistics induction is very helpful once a content management system is employed to get websites. The synchronic linguistics induction algorithm will mechanically extract generation rules, and then, a synchronic linguistics computer program will perform a reverse process to find the underlying interface linguistics.

In summary, our approach uses image analysis to acknowledge atomic interface objects and applies the SGG to specify patterns underlying websites. Supported the synchronic linguistics, a graph computer program takes a spatial graph, that is abstracted from associate interface image, as input and produces a linguistics interpretation of the interface.

IV SYSTEM IMPLEMENTATION

We have enforced a image for page segmentation. Our image primarily includes 2 subsystems: one supports image analysis to acknowledge atomic interface objects and also the other supports spatial parsing for page segmentation.

For image analysis, we have a tendency to developed the CompDetect application,which supports loading either a picture or directly from a URL. Once the image is loaded, CompDetect activates the stages of segmentation and classification. Segmentation may be done either incrementally or all promptly mechanically. The output of the segmentation stage is bestowed visually on the screen to allow manual review. The classification stage may be divided into three stages. First, produce a info, that is either loaded from a antecedently created file or created by incrementally fixing initial classification results. The second stage permits the user to classify interface objects. Once again, the user will supervise the classification result and fix them manually. The last stage permits the user to save lots of the classified objects into an XML go into order to represent an input for page segmentation in the next step.Fig.2 shows the type editor and grammar editor.

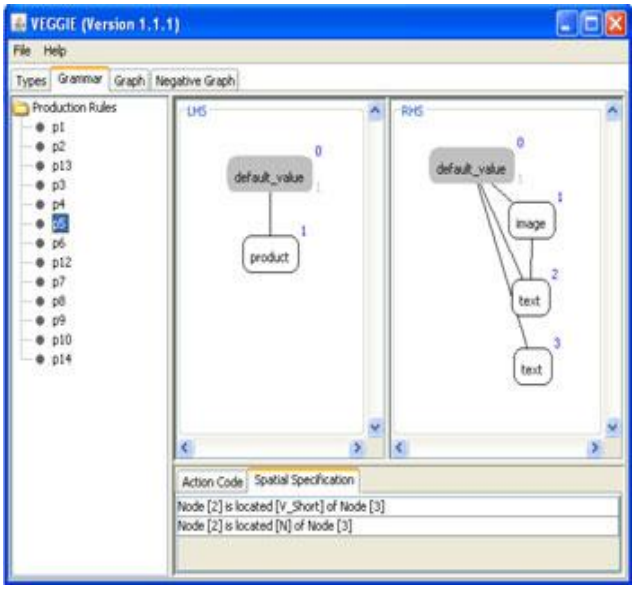

Fig.2 Grammar Editor in VEGGIE

After image analysis, the mechanically generated spatial graph is shipped to a visible programming setting, i.e., Visual Environment for Graph Grammars: Induction and Engineering (VEGGIE) [4], [5], for page segmentation. produce supports the SGG specification and parsing. produce primarily consists of three freelance editors (i.e., the sort Editor, the descriptive linguistics Editor, and also the Graph Editor) and an SGG program. The three editors give GUIs for styleers to visually design a graph grammar and square measure seamlessly operating along in produce.

Fig.3 Web interface and it's semantic interpretation

The combined views ease the change between totally different editors with the same look and feel, which boosts a coherent understanding. Descriptive linguistics designers will visually produce visual objects, i.e., node types, within the sort Editor, or import existing node sorts from a go into the shape of GraphML. Then, based on these outlined nodes, the designer will outline productions within the Grammar Editor. Within the Graph Editor, the designer will visually draw or import a spatial graph to be analyzed by the SGG program. The data files storing nodes, grammar, and graph square measure shared and interoperated by

all editors. Fig.3 shows web interface and its interpretation results.

V CONCLUSION

In the net interface interpretation, it's difficult to find the linguistics structure underlying an online interface. Different from existing heuristic approaches, this paper develops a novel approach for page segmentation supported image analysis and graph grammar induction algorithm. Rather than analyzing DOM structures, our approach uses advanced image process to acknowledge atomic interface objects associate degreed divide an interface into many regions. The image analysis produces a spatial graph, within which nodes represent recognized interface objects and edges model spatial relations, that imply a detailed semantic relation. Supported the spatial graph, a spatial parsing is performed to recover the semantics of the corresponding Web page. Owing to the distinct capability of spatial specifications within the abstract syntax, the SGGI is chosen because the definition formalism for page segmentation. We have tested our approach on the marskandspencer website, which shows promising results. As the future work, we'll conduct additional experiments to research issues like generality and potency. Especially, we will investigate the way to improve the potency of coming up with a graph descriptive linguistics. Though utilize of patterns across totally different Web sites will scale back the efforts of coming up with a graph grammar, we tend to attempt to more improve the potency by applying a descriptive linguistics-induction algorithmic rule to semi automate the grammar design.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their insightful and constructive comments that have helped us to significantly improve the presentation.

REFERENCES

1. H. Ahmadi and J. Kong, “Efficient web browsing on small screens,”in Proc. ACM Int. Working Conf. Adv. Visual Interfaces, 2008, pp. 23–30.
2. F. Ashraf, T. Ozyer, and R. Alhajj, “Employing clustering techniques for automatic information extraction from HTML documents,” IEEE Trans. Syst., Man, Cybern.—Part C: Appl. Rev., vol. 38, no. 5, pp. 660–673, Sep.2008.
3. F. Ashraf and R. Alhajjt, “ClusTex: Information extraction from HTML pages,” in Proc. 21st Int. Conf. Adv. Inf. Netw. Appl. Workshops, May 2007, pp. 355–360.
4. K. Ates, J. Kukluk, L. Holder, D. Cook, and K. Zhang, “Graph grammar induction on structural data for visual programming,” in Proc. IEEE 18th Int. Conf. Tools Artif. Intell., Nov. 2006, pp. 232–242.
5. K. Ates and K. Zhang, “Constructing VEGGIE: Machine learning for context-sensitive graph grammars,” in Proc. IEEE 19th Int. Conf. Tools Artif. Intell., Oct. 2007, pp. 456–463.
6. S. Baluja, “Browsing on small screens: Recasting web-page segmentation into an efficient machine learning framework,” in Proc. World Wide Web, 2006, pp. 33–42.
7. O. Boiman, E. Shechtman, and M. Irani, “In defense of nearest-neighbor based image classification,” in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Jun. 2008, pp. 1–9.
8. Y. Borodin, J. Mahmud, and I. V. Ramakrishnan, “Context browsing with mobiles—When less is more,” in Proc. 5th Int. Conf. Mobile Syst., Appl. Services, 2007, pp. 3–15.
9. R. Burget, “Visual HTML document modeling for information extraction,”in Proc. Reconfigurable Architectures Workshop, 2005, pp. 17–24.
10. J. Kong, K. Zhang, and X. Q. Zeng, “Spatial graph grammar for graphic user interfaces,” ACM Trans. Human-Comput. Interaction, vol. 13, no. 2, pp. 268–307, 2006.
11. N. Milic-Frayling and R. Sommerer, “SmartView: Flexible viewing ofweb page contents,” presented at the Proc. of the 11th World Wide Web Conf. (poster paper), New York, 2002.
12. S. Zheng, R. Song, and J. Wen, “Template-independent news extraction based on visual consistency,” in Proc. 22nd Nat. Conf. Artif. Intell., 2007, vol. 2, pp. 1507–1512.
13. Opera Software ASA. (2008). [Online]. Available: http://www.opera.com/products/mobile/smallscreen.
14. Y. D. Yang and H. J. Zhang, “HTML page analysis based on visual cues,” in Proc. 6th Int. Conf. Document Analysis Recognit., 2001, pp. 859–864.
15. Y. Chen, W. Y. Ma, and H. J. Zhang, “Detecting web page structure foradaptive viewing on small form factor devices,” in Proc. WorldWide Web,2003, pp. 225–233.
16. F. Paterno and G. Zichittella, “Desktop-to-mobile web adaptation through customizable two-dimensional semantic redesign,” in Proc. 3rd Int. Conf.Human-Centered Softw. Eng., 2010, pp. 79–94.
17. D. Cai, S. Yu, J. Wen, and W. Ma, “Extracting content structure for Webpages based on visual representation,” in Proc. 5th Asia Pac. Web Conf., 2003, pp. 406–417.
18. J. Kong, K. L. Ates, K. Zhang, and Y. Gu, “Adaptive mobile interfacesthrough grammar induction,” in Proc. IEEE 20th Int. Conf. Tools Artif. Intell., 2008, pp. 133–140.
19. N. Mavridis, W. Kazmi, and P. Toulis, “Friends with faces: How social networks can enhance face recognition and vice versa,” in ComputationalSocial Networks Analysis: Trends, Tools and Research Advances.Berlin, Germany: Springer-Verlag, 2009.
20. M. Labsk´y, V. Sv´atek, O. ˇSv´ab, P. Praks, M. Kr´atk´y, and V. Sn´aˇsel, “Information extraction from HTML product catalogues: from source code and images to RDF,” in Proc. IEEE/WIC/ACMInt. Conf.Web Intell., 2005, pp. 401–404.
21. T. L. Wong and W. Lam, “Adapting web information extraction knowledge via mining site-invariant and site-dependent features,” ACM Trans.Internet Technol., vol. 7, no. 1, art no. 6, 2007.
22. X. Y. Xiao, Q. Luo, D. Hong, H. Fu, X. Xie, andW. Y. Ma, “Browsing on small displays by transforming web pages into hierarchically structured subpages,” ACM Trans. Web, vol. 3, no. 1, art no. 4, 2009.
23. G. Hattori, K. Hoashi, K. Matsumoto, and F. Sugaya, “Robust web page segmentation for mobile terminal using content-distances and page layout information,” in Proc. 16th Int. Conf. World Wide Web, 2007, pp. 361–370.
24. J. Chen and K. Xiao, “Perception-oriented online news extraction,” in Proc. 8th ACM/IEEE-CS Joint Conf. Digital Libraries, 2008, pp. 363–366.