Browsing Fatigue in Handhelds: Semantic Bookmarking Spells Relief

Saikat Mukherjee
Dept. of Computer Science
Stony Brook University
Stony Brook, NY, 11794, USA
saikat@cs.sunysb.edu

I.V. Ramakrishnan
Dept. of Computer Science
Stony Brook University
Stony Brook, NY, 11794, USA
ram@cs.sunysb.edu

ABSTRACT

Focused Web browsing activities such as periodically looking up headline news, weather reports, etc., which require only selective fragments of particular Web pages, can be made more efficient for users of limited-display-size handheld mobile devices by delivering only the target fragments. Semantic bookmarks provide a robust conceptual framework for recording and retrieving such targeted content not only from the specific pages used in creating the bookmarks but also from any user-specified page with similar content semantics. This paper describes a technique for realizing semantic bookmarks by coupling machine learning with Web page segmentation to create a statistical model of the bookmarked content. These models are used to identify and retrieve the bookmarked content from Web pages that share a common content domain. In contrast to ontology-based approaches where semantic bookmarks are limited to available concepts in the ontology, the learning-based approach allows users to bookmark ad-hoc personalized semantic concepts to effectively target content that fits the limited display of handhelds. User evaluation measuring the effectiveness of a prototype implementation of learning-based semantic bookmarking at reducing browsing fatigue in handhelds is provided.

Categories and Subject Descriptors

I.7.5 [**Document and Text Processing**]: Document Capture—*Document Analysis*; H.3.3 [**Information Systems**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning—*Concept Learning*

General Terms

Algorithms, Human Factors

Keywords

Semantic Bookmarking, Handheld Device Content Adaptation, Web page partitioning

1. INTRODUCTION

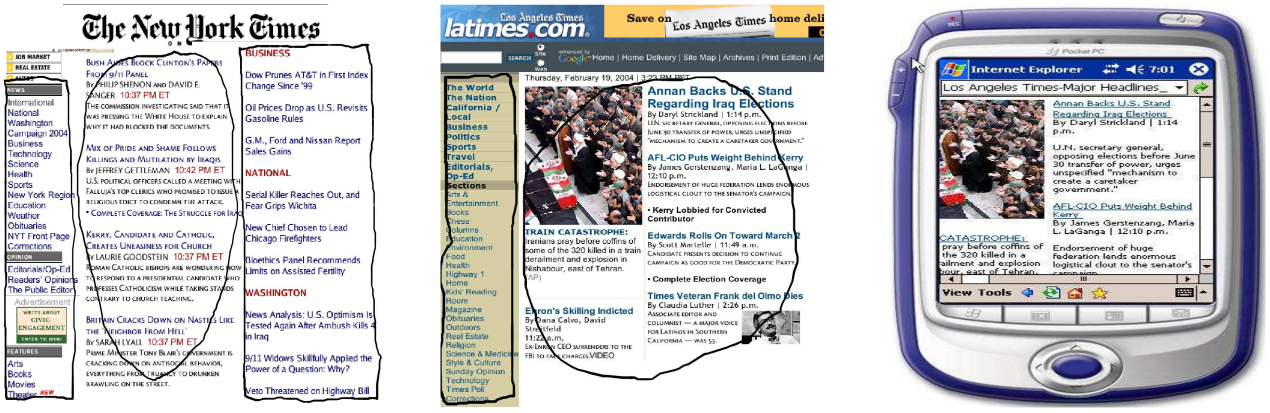
Handheld mobile devices such as PDAs and cell phones, with browsers and processors embedded in them, are becoming popular as Web browsing gadgets "on-the-go". However, their limited display size forces users to scroll tediously using various buttons to view the desired content. This makes browsing with handhelds a tedious and fatigue-inducing task. Hence, adapting Web content so as to make browsing with handhelds more efficient is an important problem that has been drawing serious research attention.

Initial approaches to adapting Web content onto handhelds [5, 21, 23] placed the burden on content providers to script Web pages specifically for such limited display devices. More recent techniques [7, 9, 12, 35] propose heuristics for adapting the content of the entire Web page into hierarchical structures summarizing the content. While they are quite effective for exploratory browsing, there are many scenarios where the user repeatedly needs targeted data from specific Web sites. Such periodic revisits usually signify the user's interest in certain specific content in these pages – e.g. the user may periodically browse news portals to read breaking news. In such situations, adapting the content of the entire Web page will require the user to repeatedly and needlessly navigate the summary structure. On the other hand delivering focused content constituting only the desired fragment of an entire page to handhelds obviates the need for needless scrolling thereby reducing stress and fatigue.

Bookmarks provide the user with direct access to pages containing specific, highly targeted content of interest. Traditionally, creating a bookmark amounts to saving the URL of the page while retrieval fetches the entire page. However, for adapting this operational aspect of bookmarks to handhelds with limited display one has to focus exclusively on the target content. This requires associating with the bookmark both the URL of the page as well as extraction expressions that when applied to the page will retrieve the desired content. In fact, research in wrapper-based data extraction techniques [24] have focused on building such expressions using various *syntactic* cues surrounding the target content in a page. However, wrappers are learned per page and are also brittle to structural variations in the page. Thus, they are not only difficult to scale across pages but are also hard to maintain over time.

We can overcome the above limitations using the notion of *semantic bookmarks*. A semantic bookmark associates content segments in Web pages, even from different Web sites, with a "concept" from an application domain. Informally, a concept represents an abstract entity that is associated with some properties. For example, the news domain will consist of concepts such as *Taxonomy* news, *Major Headline* news,

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-046-9/05/0005.



The New York Times

JOB MARKET
REAL ESTATE
AUTOS

NEWS
International
National
Washington
Campaign 2004
Business
Technology
Science
Health
Sports
New York Region
Education
Weather
Obituaries
NYT Front Page
Corrections

OPINION
Editorials/Op-Ed
Readers' Opinions
The Public Editor

Advertisement
WRITE ABOUT CIVIC ENGAGEMENT ENTER TO WIN!

FEATURES
Arts
Books
Movies
Theater NEW

Bush Aides Block Clinton's Papers From 9/11 Panel
By PHILIP SHENON and DAVID E. SANGER 10:37 PM ET
The commission investigating said that it was pressing the White House to explain why it had blocked the documents.

Mix of Pride and Shame Follows Killings and Mutilation by Iraqis
By JEFFREY GETTLEMAN 10:42 PM ET
U.S. political officers called a meeting with Falluja's top clerics who promised to issue a religious edict to condemn the attack.
• Complete Coverage: The Struggle for Iraq

Kerry, Candidate and Catholic, Creates Uneasiness for Church
By LAURIE GOODSTEIN 10:37 PM ET
Roman Catholic bishops are wondering how to respond to a presidential candidate who professes Catholicism while taking stands contrary to church teaching.

Britain Cracks Down on Nasties Like the 'Neighbor From Hell'
By SARAH LYALL 10:37 PM ET
Prime Minister Tony Blair's government is cracking down on antisocial behavior, everything from truancy to drunken brawling on the street.

BUSINESS
Dow Prunes AT&T in First Index Change Since '99
Oil Prices Drop as U.S. Revisits Gasoline Rules
G.M., Ford and Nissan Report Sales Gains

NATIONAL
Serial Killer Reaches Out, and Fear Grips Wichita
New Chief Chosen to Lead Chicago Firefighters
Bioethics Panel Recommends Limits on Assisted Fertility

WASHINGTON
News Analysis: U.S. Optimism Is Tested Again After Ambush Kills 4 in Iraq
9/11 Widows Skillfully Applied the Power of a Question: Why?
Veto Threatened on Highway Bill

(a)

Los Angeles Times latimes.com
Save on Los Angeles Times home deli

SEARCH Site Web enhanced by Google Home | Home Delivery | Site Map | Archives | Print Edition | Ad

The World
The Nation
California / Local
Business
Politics
Sports
Travel
Editorials, Op-Ed
Sections
Arts & Entertainment
Books
Chess
Columns
Education
Environment
Food
Health
Highway 1
Home
Kids' Reading Room
Magazine
Obituaries
Outdoors
Real Estate
Religion
Science & Medicine
Style & Culture
Sunday Opinion
Technology
Times Poll
Corrections

Thursday, February 19, 2004 | 3:23 PM PST

TRAIN CATASTROPHE: Iranians pray before coffins of some of the 320 killed in a train derailment and explosion in Nishabour, east of Tehran. (AP)

Enron's Skilling Indicted
By Dana Calvo, David Streitfeld
11:22 a.m.
Ex-Enron CEO surrenders to the FBI to face charges. VIDEO

Annan Backs U.S. Stand Regarding Iraq Elections
By Daryl Strickland | 1:14 p.m.
U.N. secretary general, opposing elections before June 30 transfer of power, urges unspecified "mechanism to create a caretaker government."

AFL-CIO Puts Weight Behind Kerry
By James Gerstenzang, Maria L. LaGanga | 12:10 p.m.
Endorsement of huge federation lends enormous logistical clout to the senator's campaign.

• Kerry Lobbied for Convicted Contributor

Edwards Rolls On Toward March 2
By Scott Martelle | 11:49 a.m.
Candidate presents decision to continue campaign as good for the Democratic Party.

• Complete Election Coverage

Times Veteran Frank del Olmo Dies
By Claudia Luther | 2:26 p.m.
Associate editor and columnist — a major voice for Latinos in Southern California — was 55.

(b)

Pocket PC
Internet Explorer 7:01
Los Angeles Times-Major Headlines_

Annan Backs U.S. Stand Regarding Iraq Elections
By Daryl Strickland | 1:14 p.m.

U.N. secretary general, opposing elections before June 30 transfer of power, urges unspecified "mechanism to create a caretaker government."

AFL-CIO Puts Weight Behind Kerry
By James Gerstenzang, Maria L. LaGanga | 12:10 p.m.

Endorsement of huge federation lends enormous logistical clout to the senator's campaign.

CATASTROPHE:
pray before coffins of the 320 killed in a ailment and explosion bour, east of Tehran.

View Tools

(c)

Figure 1: (a) New York Times front page (b) Los Angeles Times front page (c) Los Angeles Major Headlines instance on a PocketPC Emulator

Category news, etc. As far as properties go, *Major Headline* news items are characterized by a link labeled with the headline text, the news source, and a brief summary. Occurrence of a concept in a page is said to be its instance. In Figures 1(a) and (b) the rectangular portions on the leftmost columns are instances of *Taxonomy* news while the elliptical portions are *Major Headline* news instances and the rectangular portion on the rightmost column in Figure 1(a) is a *Category* news instance. For the end user, creating a semantic bookmark amounts to merely highlighting (some) concept instances in (a few) Web pages. Retrieval of a semantic bookmark, on the other hand, means not only extracting the concept instances from the Web pages used to create it but also from any page in any other site (specified by the user) where the concept can occur. For example, if the user creates the semantic bookmark of *Major Headline* news from the front page of New York Times then it should be possible to retrieve headline news items from Los Angeles Times front page also using this bookmark even though Los Angeles Times was not used for creating the bookmark. Observe that in contrast to a wrapper the scope of a semantic bookmark extends to all those pages across sites with similar content semantics, i.e. it is scalable.

In this paper we explore the idea of realizing semantic bookmarks by judiciously combining machine learning with Web page segmentation. Broadly speaking the method is this: Organizing a Web page into its logical structure amounts to creating a tree of partitions each of which aggregates items in the page with similar content semantics. The user highlights a (small) set of example partitions, possibly from different partition trees, as instances of the concept to be bookmarked. From these labeled nodes a statistical model of the features in the bookmarked content are learned. The learned concept models are then applied to identify and retrieve concept instances from any other partition tree that shares a common content domain (e.g. news portals, travel portals, etc.) and delivered to the handheld. Figure 1(c) shows the major headlines new fragment of the Los Angeles Times front page on a PocketPC handheld.

An alternative implementation of semantic bookmarking is through *ontologies*, a computational vehicle for capturing machine processable knowledge about an application domain. Specifically, this knowledge is represented explicitly in the ontology as domain concepts, their features, and relationships among them. In this approach, the ontology will identify the concept instances present in the page which can then be saved as semantic bookmarks. The idea of using ontologies for implementing semantic bookmarking was mentioned in [17] and [27] within the larger context of creating a semantic layer over Web pages and for assistive browsing respectively. However, content delivery to handhelds was not the focus of those works. The problem with ontology-based approaches is that it limits semantic bookmarking to concepts present a priori in the ontology. Since an ontology may not necessarily be extensive, concepts that a user is interested in capturing may not be present in the ontology. Our learning-based approach presents a more flexible paradigm where ad-hoc personalized semantic concepts can be defined and bookmarked by users. And of course learning-based semantic bookmarking of ad-hoc concepts can also nicely complement ontology-based approaches.

The rest of the paper is organized as follows. In Section 2, we describe machine learning based techniques for creating and retrieving semantic bookmarks. Section 3 contains user evaluation based on the implementation of a prototype system. It measures the effectiveness of semantic bookmarking on reducing fatigue induced by browsing using handheld devices. Sections 4 and 5 contain related work and discussions respectively.

2. LEARNING SEMANTIC BOOKMARKS

Our approach to learning semantic bookmarks rests on two processes: (i) inferring the logical structure of the Web page via structural analysis of its content, and (ii) learning the salient features present in the content of the partitions in the logical structure to build a statistical model of the concept to be bookmarked. The learned statistical model is then used for retrieving instances of the bookmarked concept. In our earlier works in [29] and [28] respectively, we had proposed a structural analysis algorithm for partitioning Web pages and a technique for learning features to annotate Web pages. In this paper we develop a computational framework for semantic bookmarking using handhelds by tightly integrating the techniques in these two works. We will briefly review the ideas underlying them in this Section.

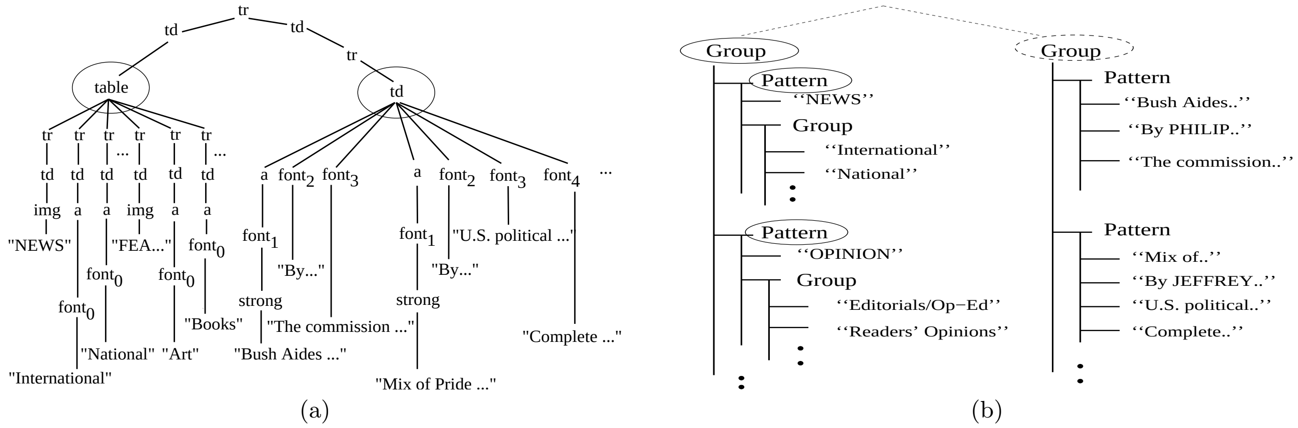


Figure 2: (a) DOM fragment of the New York Times home page (b) Partition tree of the corresponding fragment

2.1 Structural Analysis

The essence of our partitioning idea is that consistency in presentation style and spatial locality of semantically related items in Web pages can be exploited to discover sequential patterns in the DOM structure of a page. We have used a simple typing system for nodes in the DOM tree to capture these sequential patterns.

A *primitive* type encodes the presentation style (including visual cues such as font type and size) of a piece of text that corresponds to a leaf node in a DOM tree. The type of a leaf node is the sequence of HTML tags, with their attribute values, on the path from the root of the DOM tree to the node. For example, in the DOM tree of Figure 2(a) (which corresponds to the taxonomy and major headlines fragments of Figure 1(a)), all the leaf nodes corresponding to the main taxonomic items, "NEWS", "OPINION", "FEATURES", ..., etc., have the same primitive type, $tr{\cdot}td{\cdot}table{\cdot}tr{\cdot}td{\cdot}img$. Let us denote this type as T_1. Further, observe that all the subtaxonomic items, such as "International", "National", ..., etc., under each main taxonomic item, such as "NEWS", have the same primitive type, $tr \cdot td \cdot table \cdot tr \cdot td \cdot a \cdot font_0$.[1] Let us denote it using T_2.

A *compound* type summarizes the structural recurrence information at a subtree rooted at an internal node. Note that in Figure 2(a) the subtree rooted at the *table* node (shown in circle) groups together several main taxonomic items each of which is followed by a number of subtaxonomic items, *i.e.*, the entire taxonomy is clustered under this *single* DOM tree node. This property of spatial locality combined with consistency in presentation style reveals structural recurrence information about semantically related items. Observe that the sequence of primitive types of the leaf nodes in the subtree rooted at *table* is: $T_1T_2T_2 \ldots T_1T_2T_2 \ldots$. In this string the *sequential pattern*, $T_1T_2^*$ (here $*$ denotes Kleene closure), exactly captures the structural recurrence information of each semantically related item (*i.e.*, a main taxonomic item followed by a number of subtaxonomic items). Thus, the pattern $T_1T_2^*$ becomes the compound type of this *table* node.

Therefore, as illustrated by the example above, the idea underlying structural analysis is to discover sequential patterns on the typed sequence of nodes in a DOM tree. Given any two types as defined above, their equivalence is defined straightforwardly: two types are equivalent if and only if they are syntactically the same. Our structural analysis algorithm is built on the notion of *maximal repeating substrings* which is the smallest repeating substring with maximal coverage in the original string.

Since semantically related items exhibit spatial locality, structural analysis can be performed recursively bottom-up starting from the leaf nodes of the DOM tree of a HTML document. First, primitive types are assigned to all leaf nodes. The type of an internal node with only one child node is the same as that of the child. For internal nodes with multiple children, the type of each child is first computed and then the sequence of types belonging to all the children are analyzed for patterns.

Analysis of type sequences for pattern detection is an iterative process. In the first step, consecutive nodes having equivalent types are collapsed into a single node. The intuition behind this is that they all relate to the same item. We denote this node as a *group* node. The type of this group node is the same as any of it's child. Next, the modified sequence is analyzed for maximal repeating substrings. Every sequence of consecutive nodes whose types match the maximal repeat are collapsed under single nodes. These nodes are denoted as *pattern* nodes. The type of this pattern node is the sequence of types in the repeat. This procedure of grouping and pattern mining is repeated until no more patterns can be detected. If the iterations do not terminate in a single group node then the remaining non-pattern nodes are merged with their preceding pattern nodes to create a set of pattern nodes below a group node.

We illustrate pattern detection using an example type sequence $T_1T_2T_3T_2T_3T_4T_1T_2T_3T_5$. Observe that T_2T_3 is a maximal repeating substring. Let us use a new type T_6 to denote the pattern T_2T_3. Then after the first iteration, the type sequence becomes $T_1T_6T_6T_4T_1T_6T_5$. The first two occurrences of T_6 can be collapsed into a group node, resulting in $T_1T_6T_4T_1T_6T_5$, in which T_1T_6 is a maximal repeating substring. Again, we use a new type T_7 to represent the pattern T_1T6. So after the second iteration the type sequence becomes $T_7T_4T_7T_5$. No more patterns can be detected and

[1] The *font* tags with different subscripts in Figure 2(a) (*e.g.*, $font_0$) denote *font* tags with different attributes such as type and size.

the iterations stop. Finally, T_4 is merged with it's preceding pattern node T_7 while T_5 with it's preceding T_7. T_7 is the type assigned to the ultimate group node.

Figure 2(b) shows the result of our partitioning technique on the DOM fragment of New York Times in Figure 2(a). Intuitively, a *group* node in the partition tree aggregates repeated occurrences of items that are semantically similar while a *pattern* node encapsulates each such item. For instance, in Figure 2(b), the dotted circled group node aggregates occurrences of *Major Headlines* news concept. The pattern nodes below this group node correspond to every individual *Major Headlines* news item.

2.2 Concept Model

The statistical model of a concept is developed from *features* learned from the set of partition tree nodes which are labeled as its instances. Given any partition tree feature learning generates a set of features, with corresponding weights, at every node in the tree. During training, the probability of occurrence of a feature in a concept is computed by a simple frequency counting and smoothing based maximum likelihood approach.

The content of a partition as well as the style with which the content is presented are both utilized to learn *unstructured* and *structured* features. However, our learning-based framework is quite general and other kinds of features can be accommodated in it.

Unstructured Features: After eliminating stop-words the bag of words in the partition tree constitute the unstructured elements in the feature space. Each feature element is assigned a weight at every node in the partition tree.

At a leaf node p_i of a partition tree, the weight of a feature is the number of its occurrences in the text of p_i. The weight of a feature at an internal partition tree node p_i is the sum of its weights from the immediate children nodes of p_i. In this way, weights of features are propagated bottom-up the tree.

However, sometimes it is necessary to utilize the partition tree structure even further to assign higher weights to more informative features. It is often the case that Web page designers group together related content under certain words (*e.g.*, "BUSINESS," groups together the articles "Dow Prunes ..", "Oil Prices ..", and "G.M., .." in Figure 1(a)). We should assign a relatively higher weight to such words since they are in some sense the "constant" features of the content. When constructing the partition tree the non-constant items become children of a group node $p_{i'}$ and the constant item $p_{i''}$ becomes the sibling of $p_{i'}$. Together they appear as the children of a pattern node p_i. (See illustration of this process in Figure 2(b) for taxonomy news). Under these circumstances, the weights of features in $p_{i''}$ are multiplied by a factor equal to the number of children in $p_{i'}$. For instance, in the partition tree corresponding to the page in Figure 1(a), the weight of the feature "BUSINESS" will be increased by the number of children in its sibling group node (3 in this case). Subsequently, bottom-up aggregation is performed as described before.

Structured Features: Whereas unstructured features represent important words that appear in the textual content of partitions, structured features capture the presentational aspects of their content. For instance, in Figure 1(a), each *Major Headline* news item is presented as a link ("Bush Aides.."), followed by two consecutive text strings ("By..", "The commission.."). Some news items also include an optional link (e.g. "Complete.." in news item 2). Abstractly speaking the presentation style is captured by the sequence: $link \cdot text \cdot text \cdot ?link$ where $?link$ means that this $link$ may not always be present in all headline news items (akin to the ? operator used in the language of regular expressions).

The structured feature of a leaf node is either a *link* or *text* since leaf nodes in the partition tree contain either hyperlinks or text strings.[2] Hyperlink leaf nodes have only a *link* feature with a weight of 1 while text leaf nodes have only a *text* feature with unit weight. We propagate the structured features of the leaf nodes up the tree to construct the structured features of the internal nodes and assign weights to them. The structured features of internal nodes are constructed thus: If an internal node is not a pattern node then its structured feature set is just the union of its children's structured features. The weight of each feature in this set is the cumulative sum of the feature's weight in each of the node's children. Besides this unioned set of features, each pattern node also has an additional feature which reflects the repetitive structure associated with it. The repetitive structure of a pattern node is captured in this additional feature by concatenating the structured features of the node's children. Since we want to make a determination of concept instances using features that will always be present, features representing the optional aspect of the pattern are omitted. The synthesized concatenated feature is assigned a weight of 1 at the pattern node.

For instance, in Figure 2(b), the leaf partitions "Bush Aides..", "By PHILIP..", and "The commission.." have structured features *link*, *text*, and *text* respectively. Similarly, the leaf partitions "Mix of..", "By JEFFREY..", "U.S. political..", and "Complete.." have the features *link*, *text*, *text*, and *link* respectively. Structural analysis on the entire sequence of major headlines, shown in Figure 1(a), yields the set of structured features $\{\langle link{\cdot}text{\cdot}text, 1\rangle, \langle link, 1\rangle, \langle text, 2\rangle\}$ for the first pattern node. Similarly, the second pattern node has $\{\langle link{\cdot}text{\cdot}text, 1\rangle, \langle link, 2\rangle, \langle text, 2\rangle\}$ as its set of structured features. Note the *link* element denoting "Complete .." is optional and hence is discarded from the structured feature set of the 2nd pattern node. Finally, the set of structured features at the group node (considering these two pattern nodes only) is $\{\langle link \cdot text \cdot text, 2\rangle, \langle link, 3\rangle, \langle text, 4\rangle\}$.

2.3 Concept Detection

The objective now is to use the learned model to identify concept instances in the partition tree of a new Web page. The likelihood of any node in this tree being an instance of a concept is computed using a multinomial distribution on the features at that node and probabilities of occurrences of features in the concept. However, to cope with false positives and ambiguities, we augment a simplistic likelihood-based approach with a two-step process to unambiguously identify concept instance nodes. In the first step, a set of candidate partition tree nodes for a concept is generated. In the second step, a bipartite graph based technique is used to produce a set of unambiguous $\langle \text{concept}(c), \text{node}(n)\rangle$ pairs. Each $\langle c, n\rangle$ pair means that the subtree rooted under the node n in the partition tree is an instance of the concept c.

Candidate Generation: The aggregation of semantically related items by structural analysis results in the content of

[2] In this work we do not use other leaf elements such as images, etc. in our feature space.

M1	New York Times	On what did counter-terrorism officials blame 9/11?
M2	CNN	Who is the Iraqi police brigadier general?
M3	Washington Post	Where was the Taliban suspect imprisoned?
M4	Financial Times	Why did Clarke blame Bush for dereliction of duty?
M5	Houston Chronicle	What will Texas roadsides turn into in May and why?
M6	Independent	What did the leader of the train drivers union say?
M7	Los Angeles Times	How old was Frank Del Olmo when he died?
M8	Capital Times	What did Doug Moe stumble upon?
M9	-	Summarize the Iraq war news from CNN, Los Angeles Times New York Times, and Washington Post
M10	-	What is every Major Headline on?

(a)

C1	New York Times	Is British Open being discussed in Sports?
C2	CNN	Is AT&T Wireless being discussed in Business?
C3	Washington Post	Is Martha Stewart being discussed in Business?
C4	Financial Times	Is IBM being discussed in Business?
C5	Houston Chronicle	Is Disney being discussed in Business?
C6	Independent	Is Harmison being discussed in Sports?
C7	Los Angeles Times	Is Dean being discussed in Politics?
C8	Capital Times	Who struck work?
C9	-	Summarize Baseball news from Sports in New York Times, Capital Times, and Washington Post
C10	-	Count all the articles in Category News

(b)

Table 1: (a) Major Headlines News Concept Questions (b) Category News Concept Questions

a subtree rooted at a partition tree node being: (i) "close" to the content in the subtrees rooted at its children, and (ii) "distant" from the content in the subtrees of its immediate sibling nodes. For instance, in Figure 2(b), the likelihood of the dotted group node being an instance of major headlines concept is close to its children pattern nodes while being distant from its sibling group node. To compute this we used two thresholds, t_{chld} and t_{nbr}, to define the notions "close" and "distant" respectively. A node is a candidate concept instance if and only if it's average likelihood deviation from it's siblings is greater than t_{nbr} and average likelihood deviation from it's children is less than t_{chld}.

Ambiguity Resolution: Since the same node can be a candidate for different concepts, ambiguities can arise. We represent the association between concepts and candidate nodes as a bipartite graph – the set of concepts C, and the set of candidate nodes P are the two disjoint sets of vertices in the graph. An edge between $c_i \in C$ to $p_k \in P$ is created if $p_k \in Candidate(c_i)$. The idea behind bipartite graph-based ambiguity resolution is as follows: First we form the set S_i for every concept c_i. S_i consists of nodes that only match c_i. Now pick that node p_k in S_i with the maximum likelihood value to unambiguously represent an instance of the concept c_i. We remove all the other edges from c_i to any $p_l, l \neq k$ from the graph. This computation is repeated until it is not possible to derive any more 1–1 associations between concepts and partition nodes.

3. EXPERIMENTAL RESULTS

Experimental Setup: We implemented a prototype semantic bookmarking system based on the integration of logical structure of Web pages with feature learning as described in the previous section. This prototype system was executed in a desktop environment where semantic bookmarks were learned from training Web pages and were subsequently used to retrieve concept instances from a collection of test Web pages. Each Web page, training or test, was transformed into a partition tree and features extracted from every node in the tree. Instances of concepts, which were to be bookmarked, were manually identified in the training pages and their corresponding nodes in the partition trees accordingly labeled. The features in these labeled partition nodes were used to learn the concept models. The learned models were applied on the partition trees of the training pages and likelihood values computed for every concept at every node. The t_{nbr} and t_{chld} thresholds for a concept were determined by analyzing it's likelihood values at immediate siblings and children nodes, respectively, of it's labeled partition nodes. Finally, the trained models and the computed thresholds were applied on the partition trees of test Web pages and concept instance nodes identified.

In our current prototype system, identification of concept instances in training pages is performed by manually exploring the corresponding partition trees. Unlike DOM trees, the partition tree of a Web page produced as a result of structural analysis is quite shallow. Thus, it is not very difficult for an user to navigate to the node in the logical structure whose subtree corresponds to the concept instance of interest.[3]

The objective of our experiments was to compare semantic bookmarking against normal browsing for focused content retrieval in handheld devices. To this extent, we have concentrated on a *quantitative* assessment of our semantic bookmarking technique. We measured two metrics – time and I/O gestures (pen taps) users need to complete a set of focused browsing tasks with and without semantic bookmarking. These metrics were measured in a PocketPC emulator which simulates a handheld browsing environment. In-

[3] Incorporating an user-friendly interface for giving training examples is a work in progress.

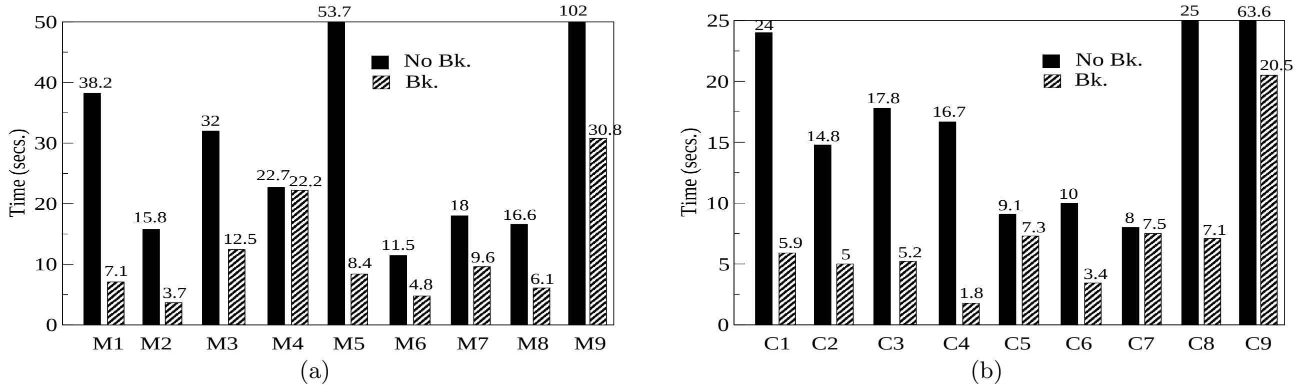


Figure 3: Time taken, with and without Semantic Bookmarks, for answering the questions in (a) Major Headlines News Concept, and (b) Category News Concept

stead of a pen or a navigation button, users perform the vertical and horizontal shift operations in the emulator browser with mouse clicks on the emulator button. Figure 1(c) shows such an emulator.

The test Web pages used in the desktop experiment were loaded up into the emulator environment. In addition the content of the concept instances identified by our learning algorithm were converted into HTML. Images present in the original Web page were preserved in the HTML conversion while scripts were removed. This HTML conversion corresponds to retrieving the semantic bookmark and rendering it on the handheld's Web browser.

Both the test Web page as well as the bookmarked content extracted from the test page were loaded into the PocketPC emulator. Evaluations were conducted on these loaded pages.

Subjects, Domains, and Tasks: We used 10 subjects as evaluators. The subjects were chosen based on their familiarity with handheld devices. Each of them had used at least one handheld device, usually a cell phone, for over a year. All the subjects were computer science graduate students who were comfortable with our test setup.

We selected the news domain and the travel domain for evaluation. These two domains possess dynamic content and are also quite popular among Web users. Prior to the experiment, the subjects were made familiar with the layout of the content in the pages chosen in the two domains. This conforms to the notion that that users bookmark content from familiar and frequently visited pages.

Subjects were given a questionnaire and their task was to answer it w.r.t the information content in test page and the bookmarked content loaded in the handheld. The tasks were divided into three categories with increasing levels of difficulty:

- Answering questions from single Web pages.
- Answering questions that require comparing information from a set of Web pages.
- Answering questions that require exhaustively reading the retrieved bookmark from all of the Web pages.

The motivation behind this gradation of tasks was to evaluate the effectiveness of semantic bookmarking for comprehending information not just from a single page but from a collection of pages in the same domain.

We used the front pages of 8 news portals as the test set for our experiments on the news domain. In each of these pages, we identified two semantic concepts *Major Headlines News* and *Category News*. The content in these concept instances are very dynamic in nature and as such are suitable to be bookmarked. Two front pages, one each from New York Times and CNN, were used for training purposes[4]. Table 1 shows the tasks for the concepts in the news domain. The first column in each concept's table corresponds to the task number, while the second column is a news site, and the third column is the question which has to be answered from the front page of that site in the test set. The first 8 tasks for both the news concepts are single page questions, while question 9 compares four Web pages, and the last question is exhaustive in nature.

The front pages of Expedia, Priceline, and Orbitz were used for evaluation in the travel domain. The semantic concept of *Travel Deals*, which shared the dynamic content nature of the news concepts, was used for bookmarking. An Expedia front page was used for training this concept. Table 3(a) shows the tasks in the travel domain for this concept. Questions $D1, D2$, and $D3$ are single page questions, while $D4$ is across pages, and answering $D5$ requires exhaustive enumeration of all the deals in all the three pages.

Each subject was required to answer all the 20 questions from the news domain as well as all the 5 questions from the travel domain. In order to smooth the effect of the order of experimentation, each of 5 randomly chosen subjects answered the questions first with and then without semantic bookmarking. The remaining 5 subjects carried out the experiments in the reverse order. Moreover, for each subject, a time gap of 7 days was observed between answering the first and second sets of questions. Since we did not discern any noticeable difference between the two groups of subjects, i.e. those who evaluated first with semantic bookmarks and those who evaluated first without semantic bookmarks, the results shown in the following subsections are averaged over all the 10 users.

Results on Time: Figures 3(a) and (b) show the time taken, averaged over all the 10 subjects, to accomplish the first nine tasks in the *Major Headlines News* and *Category News* concepts respectively. In both the figures, the shaded

[4] The pages used in the test set for these two sites were different from the training pages.

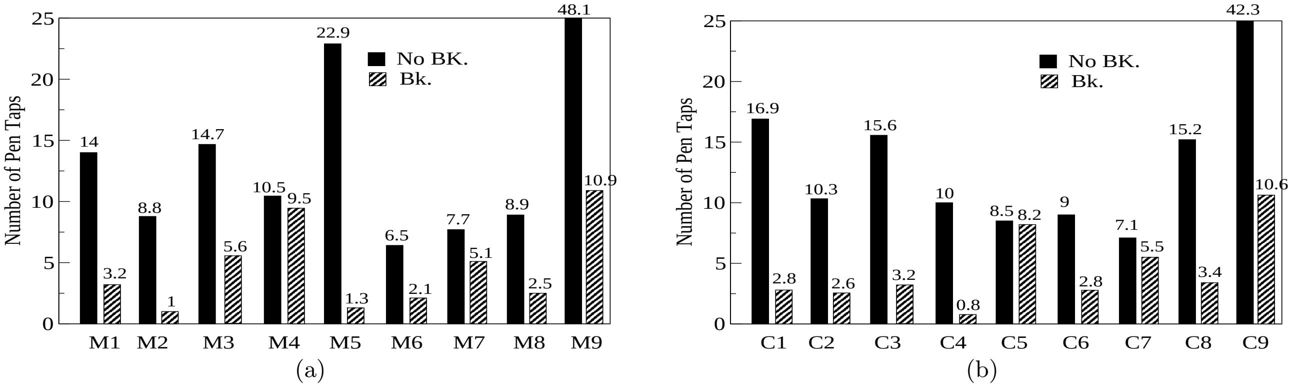


Figure 4: Number of Pen Taps required, with and without Semantic Bookmarks, for answering the questions in (a) Major Headlines News Concept, and (b) Category News Concept

bars correspond to time taken without semantic bookmarking while the checkered bars correspond to time with semantic bookmarking of the corresponding concept. The numbers do not include the time taken to load up the pages in the emulator browser since we were concerned only with comparing the information comprehension times between the two approaches. For the same reason, the numbers do not include the (insignificant) time required to compute the semantic bookmark also.

Observe the significant decrease in time with the use of semantic bookmarking for both the concepts. For the *Major Headlines News* concept this decrease ranges from 84.36% in $M5$ to 2.2% in $M4$ with an average decrease of 47.37% over the first eight tasks. In the *Category News* concept this decrease ranges from 89.22% in $C4$ to 6.25% in $C7$ with an average decrease of 46.53% over $C1$ to $C8$. For the cross page questions, $M9$ and $C9$, there are decreases of 69.80% and 67.77% in time respectively. The decrease in times, for both the concepts, varies between sites due to the difference in layout styles among them. Thus, while the layout of major headlines news in Financial Times ($M4$) facilitates easy browsing even without semantic bookmarking, the complex layout of the Houston Chronicle major headlines news ($M5$) provides evidence of the usefulness of semantic bookmarking. For most of the tasks in Figures 3(a) and (b), the *Category News* concept times are less than the corresponding times in *Major Headlines News.* This is due to the organization of category news into subcategories which makes information access easier. The time portions in Table 2 show the effect of semantic bookmarking for the exhaustive questions $M10$ and $C10$. Averaged over all the eight sites, the decreases in time are 50.05% and 41.02% for *Major Headlines News* and *Category News* respectively.

Similar decreases in time are also observed for the tasks related to the *Travel Deals* concept in the travel domain as shown in Figure 5 and Table 3(b) (time portions). The increased average decrease in time over $D1$, $D2$, and $D3$, 84.5%, compared to the news domain is due to the very complex layout of information with forms and search boxes in travel front pages.

Results on I/O: Figures 4(a) and (b) show the decrease in I/O gestures, i.e. pen taps, averaged over all the 10 subjects with the use of semantic bookmarking in the news domain. For *Major Headlines News*, this decrease ranges from 94.32% in $M5$ to 9.52% in $M7$ with an average decrease of 63.11% over the first eight tasks. Similarly, for *Category News* the decrease ranges from 92% in $C4$ to 22.53% in $C7$ with an average decrease of 62.78% over $C1$ to $C8$. The cross page questions, $M9$ and $C9$, have decreases of 77.34% and 74.94% respectively. Table 2 shows the decrease in pen taps for the exhaustive questions $M10$ and $C10$. Averaged over all the eight pages, there are decreases of 65.86% and 57.78% for $M10$ and $C10$ respectively.

The average decrease in pen taps for the *Travel Deals* concept, as shown in Figure 5, over $D1$, $D2$, and $D3$ is around 91.87%. Similar decrease in pen taps are also observed for the cross page question $D4$ and the exhaustive question $D5$ as shown in Figure 5 and Table 3(b) respectively.

Results on Bandwidth: In a mobile handheld environment, the bandwidth of the wireless network poses constraints on the amount of data that can be transmitted. Table 4 summarizes our findings on the bandwidth savings which could be accomplished by the use of semantic bookmarks. The first column in Table 4 indicates the front page of the Web site, the second column shows the total number of bytes including images, scripts, and plain HTML for that page, the third column (C_3) shows the total number of bytes without scripts, and the fourth column (C_4) shows the total number of bytes without images and scripts. The first column (C_5) in each news concept shows the number of bytes, including images but excluding scripts, for that concept instance in the corresponding Web page. The second column in each news concept shows the %age reduction of C_5 over C_3 while the third column shows the %age reduction of C_5 over C_4. Observe the significant reduction in bandwidth in most of the pages and across both the concepts even when semantic bookmarks with images is compared to original Web page without images. This indicates the utility of semantic bookmarking, from a hardware perspective, for focused repetitive browsing activities.

4. RELATED WORK

The problem of creating, retrieving and evaluating the effectiveness of personalized semantic bookmarks for handhelds is a relatively new topic in the literature. The areas closely related to this work include content adaptation for small-screen devices, wrappers for data extraction, and the Semantic Web.

Initial efforts at adapting Web content onto handhelds re-

Web Page	Major Headlines						Category					
	Pen Taps			Time (secs)			Pen Taps			Time (secs)		
	No Bk.	Bk.	% Red.	No Bk.	Bk.	% Red.	No Bk.	Bk.	% Red.	No Bk.	Bk.	% Red.
New York Times	16.2	4.0	75.31	32.5	12.9	60.31	29.4	14.6	50.34	54.5	35.9	34.13
CNN	7.7	2.0	73.92	10.3	5.0	51.46	16.7	5.4	67.66	28.6	17.6	38.46
Washington Post	20.0	9.4	53.00	39.1	22.6	42.20	19.1	6.0	68.59	40.1	17.1	57.36
Financial Times	17.3	8.6	50.29	41.3	23.1	44.07	15.1	3.8	74.83	26.8	12.2	54.48
Houston Chronicle	25.1	3.9	84.46	50.7	23.9	52.86	14.3	10.2	28.67	31.6	25.0	20.89
Independent	10.4	4.0	61.54	19.0	9.1	52.11	14.0	4.3	69.29	21.8	12.8	41.28
Los Angeles Times	18.6	9.9	46.77	30.8	19.4	37.01	24.7	11.5	53.44	45.6	26.2	42.54
Capital Times	21.7	4.0	81.57	27.5	10.9	60.36	17.8	9.0	49.44	21.8	13.3	38.99

Table 2: Exhaustive Question (M10 and C10) for News Domain Concepts

D1	Expedia	Is there a deal to Florida?
D2	Orbitz	Is there a deal to Florida?
D3	Priceline	Is there a deal to Florida?
D4	-	What is the cheapest deal to Florida from Expedia, Orbitz, and Priceline?
D5	-	How many deals there are?

(a)

Web Page	Deals					
	Pen Taps			Time (secs.)		
	No Bk.	Bk.	% Red.	No Bk.	Bk.	% Red.
Expedia	16.2	5.1	68.52	30.8	14.6	52.60
Orbitz	30.1	3	90.03	41.3	17.8	56.90
Priceline	21.1	5.4	74.41	34.2	12.5	63.45

(b)

Table 3: (a) Travel Deals Concept Questions (b) Pen Taps and Time required, with and without Semantic Bookmarks, for answering Question D5

lied on WML (Wireless Markup Language) and WAP (Wireless Application Protocol) for designing and displaying Web pages [23, 5, 21]. That these approaches impose additional burden on Web page authors to create separate WML content, led to work on automatic adaptation of normal Web content onto small screen devices (see [6, 9, 8, 7, 35, 12, 38, 20, 10, 2]). These works have focused on organizing the Web page into tree structures and summarizing its content. While they are effective for ad-hoc exploratory browsing, summary structures cause needless navigational steps when a user is only interested in targeted content. Our technique only presents the desired information and our evaluation results indicate that it mitigates browsing fatigue caused by needless navigation.

Recall that our technique partitions Web pages into semantically related units prior to building the statistical model. Web page partitioning techniques have been proposed for adapting content onto small screen devices [6, 9, 8, 7, 12, 38, 4, 25]. Related partitioning techniques have also been proposed for other applications like content caching [32], Web page cleaning and data mining [36, 37, 3], Web search [39], schema extraction [13], and displaying content in a browser [26]. Unlike our approach, these works do not associate content semantics with consistency of presentation style and spatial locality – the key to inferring the logical structure of a page organized around its content semantics. Semantically related items are more accurately identified and aggregated together at various levels of granularity by content analysis based on this idea. Learning salient features of partitions constituting such aggregated items enables users to create and retrieve succinct semantic bookmarks which precisely correspond to the desired content. The idea of learning features of Web page segments was recently explored in [33]. Apart from the difference in the application scenario – data cleaning in [33] vs. our semantic bookmarking – their learning setting does not utilize the presentational aspects of the content. But the fundamental difference between our work and all the above works is that we tightly integrate the logical structure of Web pages with feature learning. It is this tight coupling that facilitates identification of the more distinguishing characteristics of concepts thereby leading to the creation and retrieval of semantic bookmarks with a high degree of precision.

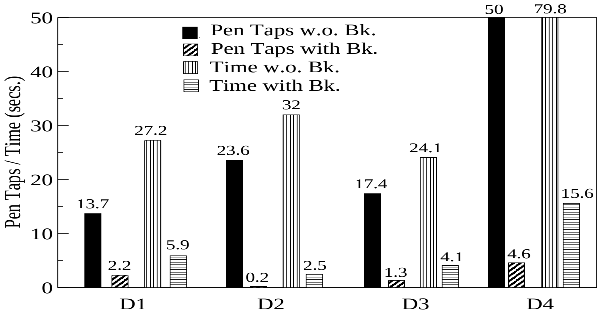


Figure 5: Pen Taps and Time required, with and without Semantic Bookmarks, for answering Questions D1 to D4

Web	Total	HTML + Img.	HTML	Major Headlines			Category		
Page	(Bytes)	(Bytes)	(Bytes)	(Bytes)	% Red.	% Red.	(Bytes)	% Red.	% Red.
New York Times	186,523	184,855	71,212	4,531	97.55	93.64	14,191	92.32	80.07
CNN	200,920	133,168	52,983	7,100	94.67	86.60	14,918	88.80	71.84
Washington Post	439,672	416,975	97,993	118,533	71.57	-	29,732	92.87	69.66
Financial Times	194,154	121,412	53,270	43,050	64.54	19.19	13,241	89.09	75.18
Houston Chronicle	227,005	186,500	69,086	34,248	81.64	50.43	9,214	95.06	86.67
Independent	85,899	72,372	25,845	2,259	96.88	91.26	4,424	93.89	82.88
Los Angeles Times	139,702	104,069	79,996	23,336	77.58	70.83	20,123	80.66	13.77
Capital Times	106,031	100,153	18,928	4,107	95.90	78.30	70,945	29.16	-

Table 4: Bandwidth Savings from Semantic Bookmarks in the News Domain Pages

Semantic bookmarking is also related to the extensive research on manually or semi-automatically constructing wrappers for data extraction from Web pages (see [24] for a survey on wrappers). However, being syntax-based, wrappers are sensitive to structural changes in the Web page. In addition, they are page-specific. Recent approaches to automated wrapper construction also rely on syntax-based solutions [1, 15, 11] (such as assuming a common schema or using specific tags as record boundary separators). In contrast, semantic bookmarking is resilient to structural changes. As long as the features associated with the bookmarked concept are sufficiently preserved in a Web page, the content corresponding to the concept instance in the page can be retrieved. Moreover, the scope of semantic bookmarking extends to pages drawn from different Web sites that share a common application domain. The notion of using "semantics" for wrapper learning was only very recently discussed in [34]. However, their use of semantics is limited to simple words and does not make use of presentational aspects of content. Moreover, unlike ours, the work in [34] does not involve inferencing of logical structures of Web pages.

The Semantic Web has spurred research on making Web pages machine understandable. To realize the Semantic Web one has to annotate Web pages with semantic meta-information. Powerful ontology management systems and knowledge bases have been used for interactive annotation of web pages [19, 22, 18] or have been combined with linguistic analysis for fully automated approaches [16, 30, 14, 29]. While ontologies and knowledge bases can be used for semantic bookmarking via Semantic Web browsers [17, 31] they however restrict the user to only those concepts defined in them. In contrast, use of machine learning facilitates creation of *personalized* ad-hoc semantic bookmarks. Such a degree of personalization not only gives users the flexibility to define their own view of semantic concepts but also provides them with a transparent workaround when a desired concept does not exist in the knowledge base.

Finally, partitioning documents into distinct segments is related to work on topic detection [40]. However, in contrast to typical topic detection works on unstructured text, our techniques analyze semi-structured HTML documents where use is made of their additional structural information.

5. DISCUSSIONS

In this paper, we have reported on a preliminary quantitative evaluation of learning-based semantic bookmarking on handheld devices. While further statistical analysis of the data is required, we believe it is also important to measure the qualitative impact of the technology on users. In particular, it would be interesting to assess user response to the loss of surrounding context versus the browsing efficiency gained by focused content delivery.

From an experimental perspective, it is worthwhile evaluating the effectiveness of semantic bookmarking on actual handhelds in a real-world wireless setting. Such a setting can give rise to additional usability issues that may not manifest themselves in an emulator environment. Implementation of semantic bookmarking in such a real environment is in progress.

Acknowledgments: This work was supported in part by the NSF grants CCR-0205376 and CCR-0311512.

6. REFERENCES

[1] A. Arasu and H. Garcia-Molina. Extracting structured data from web pages. In *ACM Conf. on Management of Data (SIGMOD)*, 2003.

[2] Y. Aridor, D. Carmel, Y. Maarek, A. Soffer, and R. Lempel. Knowledge encapsulation for focussed search from pervasive devices. In *Intl. World Wide Web Conf. (WWW)*, 2001.

[3] Z. Bar-Yossef and S. Rajagopalan. Template detection via data mining and its applications. In *Intl. World Wide Web Conf. (WWW)*, 2002.

[4] T. Bickmore and B. Schilit. Digestor: Device-independent access to the world wide web. In *Intl. World Wide Web Conf. (WWW)*, 1997.

[5] G. Buchanan, S. Farrant, M. Jones, H. Thimbleby, G. Marsden, and M. Pazzani. Improving mobile internet usability. In *Intl. World Wide Web Conf. (WWW)*, 2001.

[6] O. Buyukkoten, H. Garcia-Molina, and A. Paepcke. Focussed web searching with PDAs. In *Intl. World Wide Web Conf. (WWW)*, 2000.

[7] O. Buyukkoten, H. Garcia-Molina, and A. Paepcke. Accordion summarization for end-game browsing on

PDAs and cellular phones. In *ACM Conf. on Human Factors in Computing Systems (CHI)*, 2001.

[8] O. Buyukkoten, H. Garcia-Molina, and A. Paepcke. Seeing the whole in parts: Text summarization for web browsing on handheld devices. In *Intl. World Wide Web Conf. (WWW)*, 2001.

[9] O. Buyukkoten, H. Garcia-Molina, A. Paepcke, and T. Winograd. Power browser: Efficient web browsing for PDAs. In *ACM Conf. on Human Factors in Computing Systems (CHI)*, 2000.

[10] D. Chalmers, M. Sloman, and N. Dulay. Map adaptation for users of mobile systems. In *Intl. World Wide Web Conf. (WWW)*, 2001.

[11] C.-H. Chang and S.-C. Lui. IEPAD: Information extraction based on pattern discovery. In *Intl. World Wide Web Conf. (WWW)*, 2001.

[12] Y. Chen, W.-Y. Ma, and H.-J. Zhang. Detecting web page structure for adaptive viewing on small form factor devices. In *Intl. World Wide Web Conf. (WWW)*, 2003.

[13] C. Y. Chung, M. Gertz, and N. Sundaresan. Reverse engineering for web data: From visual to semantic structures. In *Intl. Conf. on Data Engineering (ICDE)*, 2002.

[14] P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web. In *Intl. World Wide Web Conf. (WWW)*, 2004.

[15] V. Crescenzi, G. Mecca, and P. Merialdo. Roadrunner: Towards automatic data extraction from large web sites. In *Intl. Conf. on Very Large Data Bases (VLDB)*, 2001.

[16] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. Tomlin, and J. Yien. SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. In *Intl. World Wide Web Conf. (WWW)*, 2003.

[17] M. Dzbor, J. Domingue, and E. Motta. Magpie - towards a semantic web browser. In *Intl. Semantic Web Conf. (ISWC)*, 2003.

[18] S. Handschuh and S. Staab. Authoring and annotation of web pages in CREAM. In *Intl. World Wide Web Conf. (WWW)*, 2002.

[19] J. Heflin, J. A. Hendler, and S. Luke. SHOE: A blueprint for the semantic web. In D. Fensel, J. A. Hendler, H. Lieberman, and W. Wahlster, editors, *Spinning the Semantic Web*, pages 29–63. MIT Press, 2003.

[20] M. Jones, G. Marsden, N. Mohd-Nasir, K. Boone, and G. Buchanan. Improving web interaction on small displays. In *Intl. World Wide Web Conf. (WWW)*, 1999.

[21] E. Kaasinen, M. Aaltonen, J. Kolari, S. Melakoski, and T. Laakko. Two approaches to bringing internet services to WAP devices. In *Intl. World Wide Web Conf. (WWW)*, 2000.

[22] J. Kahan, M. Koivunen, E. Prud'Hommeaux, and R. Swick. Annotea: An open rdf infrastructure for shared web annotations. In *Intl. World Wide Web Conf. (WWW)*, 2001.

[23] A. Kaikkonen and V. Roto. Navigating in a mobile XHTML application. In *ACM Conf. on Human Factors in Computing Systems (CHI)*, 2003.

[24] A. Laender, B. Ribeiro-Neto, A. da Silva, and J. Teixeira. A brief survey of web data extraction tools. *SIGMOD Record*, 31(2), 2002.

[25] W. Lum and F. Lau. A context-aware decision engine for content adaptation. *IEEE Pervasive Computing*, 1(3), 2002.

[26] N. Milic-Frayling and R. Sommerer. Smartview: Flexible viewing of web page contents. In *Intl. World Wide Web Conf. (WWW)*, 2002.

[27] S. Mukherjee, I. Ramakrishnan, and M. Kifer. Semantic bookmarking for non-visual web access. In *ACM Conf. on Assistive Technologies (ASSETS)*, 2004.

[28] S. Mukherjee, I. Ramakrishnan, and A. Singh. Bootstrapping semantic annotation for content-rich html documents. In *Intl. Conf. on Data Engineering (ICDE)*, 2005.

[29] S. Mukherjee, G. Yang, and I. Ramakrishnan. Automatic annotation of content-rich html documents: Structural and semantic analysis. In *Intl. Semantic Web Conf. (ISWC)*, 2003.

[30] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. KIM - semantic annotation platform. In *Intl. Semantic Web Conf. (ISWC)*, 2003.

[31] D. Quan and D. Karger. How to make a semantic web browser. In *Intl. World Wide Web Conf. (WWW)*, 2004.

[32] L. Ramaswamy, A. Iyengar, L. Liu, and F. Douglis. Automatic detection of fragments in dynamically generated web pages. In *Intl. World Wide Web Conf. (WWW)*, 2004.

[33] R. Song, H. Liu, J.-R. Wen, and W.-Y. Ma. Learning block importance models for web pages. In *Intl. World Wide Web Conf. (WWW)*, 2004.

[34] T.-L. Wong and W. Lam. Text mining from site invariant and dependent features for information extraction knowledge adaptation. In *SIAM Intl. Conf. on Data Mining (SDM)*, 2004.

[35] C. Yang and F. L. Wang. Fractal summarization for mobile devices to access large documents on the web. In *Intl. World Wide Web Conf. (WWW)*, 2003.

[36] L. Yi and B. Liu. Eliminating noisy information in web pages for data mining. In *ACM Conf. on Knowledge Discovery and Data Mining (SIGKDD)*, 2003.

[37] L. Yi and B. Liu. Web page cleaning for web mining through feature weighting. In *Intl. Joint Conf. on Artificial Intelligence (IJCAI)*, 2003.

[38] X. Yin and W. S. Lee. Using link analysis to improve layout on mobile devices. In *Intl. World Wide Web Conf. (WWW)*, 2004.

[39] S. Yu, D. Cai, J.-R. Wen, and W.-Y. Ma. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In *Intl. World Wide Web Conf. WWW*, 2003.

[40] J. Allen, editor. Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic Publishers, 2002.