UNIVERSITÄT AUGSBURG

Progressive Content Delivery for Mobile E-Services

M. Wagner, W. Kießling, W.-T. Balke

Report 2002-3 April 2002

INSTITUT FÜR INFORMATIK

D-86135 AUGSBURG

Institut für Informatik
Universität Augsburg
D–86135 Augsburg, Germany
http://www.Informatik.Uni-Augsburg.DE

Progressive Content Delivery for Mobile E-Services

Matthias Wagner* Werner Kießling+ Wolf-Tilo Balke+

*4th Generation Middleware Laboratory
NTT DoCoMo Euro–Labs
wagner@docomolab-euro.com

+Institute of Computer Science
University of Augsburg, Germany
{kiessling, balke}@informatik.uni-augsburg.de

Abstract

Abstract: *In this paper we present a framework for the progressive delivery of Web documents in mobile Internet services. Progressive delivery enables users to get fast access to the most relevant parts of a document. Given the reduced bandwidth and the high costs of mobile communication the idea of progressive delivery offers a promising improvement especially for mobile e-services. The central part of the delivery consists of innovative concepts for content selection to determine the most relevant document parts for successive delivery maintaining the documents' readability. To make this selection as flexible and effective as possible we consider the user's notion of relevance together with semantic author annotations and structural document characteristics. Using XML technology documents are automatically adapted to fit both personal user profiles and device constraints. A prototypical mobile news service exemplifies our approach to content selection, but our framework promises to be applicable to a broad range of future Internet services.*

Keywords: Mobile Internet, Personalized Mobile Access, User Profiling, Cooperative Information Systems, Databases, Web and XML.

1 Introduction

The further improvement of efficient web-enabled multimedia database and middleware systems is a major topic in today's database research. Managing multimedia content differs essentially from handling numeric and character data in conventional systems and requires new strategies in data storage, administration and retrieval. Apart from basic data management the delivery of multimedia content is of essential importance: today value-added Internet applications like portal services, multimedia libraries or e-shops already have to provide a flexible delivery of multimedia data and complex digital content to support different kinds of Internet users and devices. With the potential convergence of Internet technologies and mobile communication the existing demand for personalization of e-services will even increase. The current trend towards a "Mobile Internet" points to a broad acceptance of flexible, adaptable and personalizable multimedia applications in public and private life within the near future.

Within the scope of the HERON project [12] and the research initiative Preference World [5] we have developed a middleware framework called Multimedia Delivery for the efficient and effective adaptation

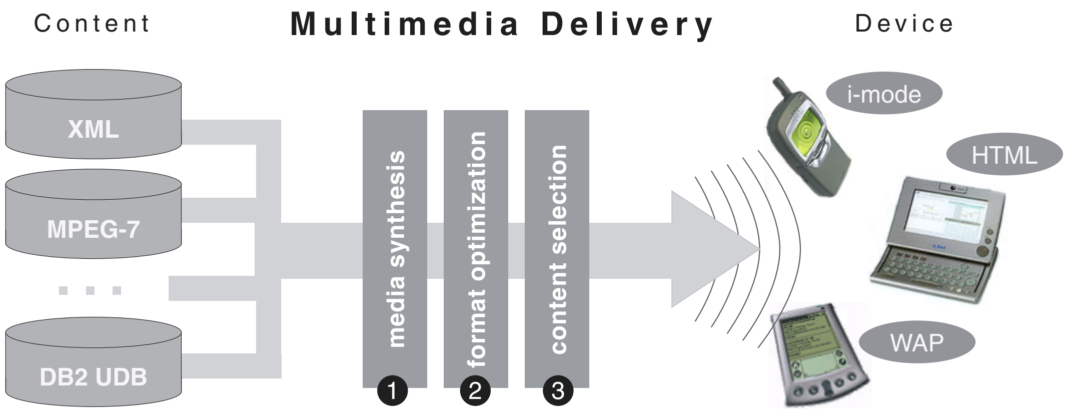


Figure 1: The conceptual layers of Multimedia Delivery.

and delivery of multimedia content to different users and mobile terminals with varying user profiles and specific device capabilities (cf. table 1). Multimedia Delivery is to strengthen the infrastructure of future Internet applications and to facilitate the implementation of multi-channel e-services for different kinds of users and different types of mobile devices. Figure 1 illustrates the conceptual layers of Multimedia Delivery and shows how the delivery framework can be deployed in practice. The building blocks of Multimedia Delivery are as follows (cf. figure 1):

❶ **Media synthesis:** Multi-channel Internet applications have to deal with media objects in varying formats. The synthesis layer of Multimedia Delivery improves access to objects through conversion and allows for the efficient computation of media objects.

❷ **Format optimization:** Multimedia database systems such as IBM DB2 Universal Database or Oracle 9*i* are frequently deployed within e-services for the storage of multimedia content. The format optimization layer is targeted at the optimal format-centered configuration of database servers supporting a cost-based and self-tuning storage optimization.

❸ **Content selection:** For final delivery to an individual user a complex multimedia document can be adapted to the user's preferences and the technical constraints of mobile terminals. To optimize the personalized delivery w.r.t. a user profile parts of a document are preselected and progressively delivered with increased priority according to their relevance to the user.

Details on the implementation and evaluation of the media synthesis and format optimization layer of our delivery framework are published elsewhere [13, 11]. In this paper we focus on the third framework layer aiming at the progressive delivery of complex multimedia documents.

The main goal of content selection is to provide an optimal delivery order for progressive delivery. Due to the possibly limited capacity in mobile environments (e.g. low bandwidths) the selection of content is twofold: with respect to a user's interests the most relevant parts of the documents should be delivered first to satisfy the user's information need. However, to maintain the readibility and semantic links within a document it is necessary to avoid a high grade of fragmentation. To identify content-bearing document parts relevance weightings are assigned to different parts of the structured document. These weightings are determined by a ranking model taking several criteria into account. Subsequently the document structure is altered such that highly weighted parts will be delivered first. Figure 2 depicts the single steps of

Device	NEC n21i	Palm V	Siemens SL45i
Markup Language	i-mode	HTML	WAP
Display Size	120x160	160x160	85x50
Image Formats	WBMP, GIF, JPEG	GIF, JPEG, PNG	WBMP
Color Depth	8 Bit	16 Bit	2 Bit
Connection	GPRS	GPRS, WLAN	GSM

Table 1: Specific device constraints of different mobile terminals.

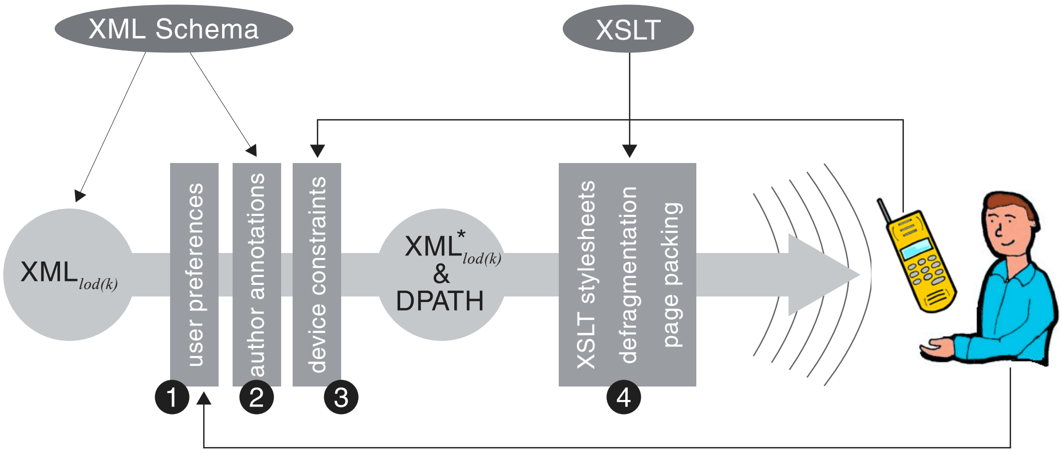


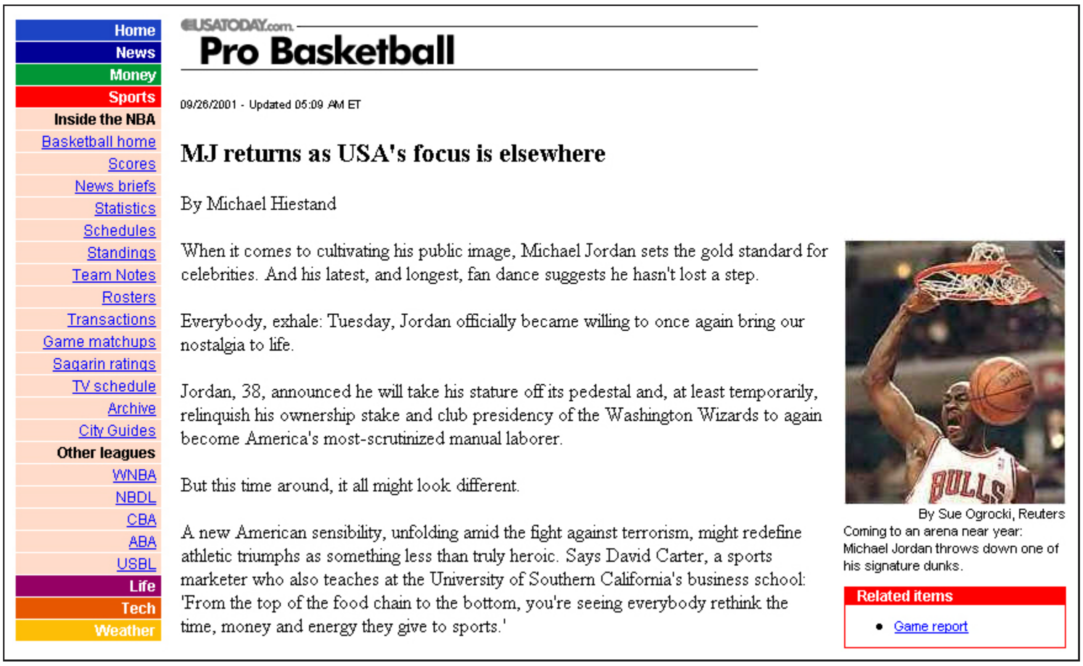
Figure 2: Stepwise selection of content and progressive delivery.

Content Selection for the delivery of structured documents to a mobile client. In this framework structured documents are processed and delivered regarding the personalized relevance of textual information, technical constraints of the (mobile) client device as well as the intended document structure and the author's semantic knowledge.

Content Selection is targeted at the successive delivery of document nodes in the best matching order for individual users. Therefore the document structure is annotated using its type definition by the author or editor of the document. In addition the degree of fragmentation in the resulting target document is lowered by heuristic algorithms regarding overall structural characteristics of the content. The extensible markup language (XML) [2] is well suited for the description of compound documents and for flexible presentation of content. Using XML technology to implement a personalizable E-Service for the progressive delivery of news articles over the Internet we will exemplify all delivery steps: Section 2 introduces our prototypical news service implementing the content selection framework. Section 3 shows how to obtain the document's readibility throughout the entire progressive delivery. Section 4 deals with related delivery approaches. Finally we close with a short summary and an outlook.

2 A mobile Internet news service

We will now consider an application that will help us to understand all concepts of content selection. We will apply the techniques described above in the framework of a personalized Internet news service with delivery to various mobile client devices. Our e-service will process and deliver newspaper articles that



Home
News
Money
Sports
Inside the NBA
Basketball home
Scores
News briefs
Statistics
Schedules
Standings
Team Notes
Rosters
Transactions
Game matchups
Sagarin ratings
TV schedule
Archive
City Guides
Other leagues
WNBA
NBDL
CBA
ABA
USBL
Life
Tech
Weather

USATODAY.com

Pro Basketball

09/26/2001 - Updated 05:09 AM ET

MJ returns as USA's focus is elsewhere

By Michael Hiestand

When it comes to cultivating his public image, Michael Jordan sets the gold standard for celebrities. And his latest, and longest, fan dance suggests he hasn't lost a step.

Everybody, exhale: Tuesday, Jordan officially became willing to once again bring our nostalgia to life.

Jordan, 38, announced he will take his stature off its pedestal and, at least temporarily, relinquish his ownership stake and club presidency of the Washington Wizards to again become America's most-scrutinized manual laborer.

But this time around, it all might look different.

A new American sensibility, unfolding amid the fight against terrorism, might redefine athletic triumphs as something less than truly heroic. Says David Carter, a sports marketer who also teaches at the University of Southern California's business school: 'From the top of the food chain to the bottom, you're seeing everybody rethink the time, money and energy they give to sports.'

By Sue Ogrocki, Reuters
Coming to an arena near year: Michael Jordan throws down one of his signature dunks.

Related items
- Game report

Figure 3: Sample news article as displayed on the Web.

are taken from various on-line newspapers, e.g. USA Today News & Information [10]. All articles are stored using a generic and device independent XML format. Figure 3 shows a sample article from USA Today's sports archive (September 26, 2001) and its XML representation (figure 4). The header of our sample document contains several XML tags for the encoding of news articles, such as the documents title, its author and posting date, etc. The document body itself contains a photo followed by two text sections as the major textual components.

2.1 Level of detail

For delivery the complex XML document has to be broken down to single blocks with a certain level of detail (*lod*). These blocks form the smallest units of information and must be delivered without further division. By definition XML documents form tree-like structures. Basically these documents consist of markup nodes, which are assembled inductively to document trees.

Figure 5 depicts the tree structure corresponding to our sample news article from figure 3 and figure 4, respectively. For simplicity we refer to a document as a whole by its root node, e.g. n_1 represents the complete sample article, whereas n_{12} refers to a partial document – the first section of text. This figure also illustrates how a document is split into a hierarchy of *lod*s, which includes a hierarchy of the XML nodes used to store information: For our sample document n_1 the highest *lod* is 2 including just the root node n_1 itself. On a lower detail level the document is partitioned into several sub-documents with each partial document being evaluated and ranked for progressive delivery, e.g. at level 1 relevance weights will be assigned to the sub-documents n_2, n_6, n_{12}, n_{16}.

Note that the different levels of a document induce an order on the XML tags used to markup its content. In our running example the tags used to markup the "title" as part of a "headline" within a "news" article are ordered as follows: news $>$ headline $>$ title. Whereas the tags used to markup a standard text "par"

```xml
<?xml version='1.0'?>
<news>
  <headline>
    <posted date='Wednesday, September 26, 2001, 05:09 AM ET'/>
    <title>MJ returns as USA's focus is elsewhere</title>
    <author>Michael Hiestand</author>
  </headline>
  <photo>
    <image type="JPEG" resolution="75" origin="/multimedia/MJreturns.jpg"/>
    <caption>Coming to an arena near year: Michael Jordan throws down one of his
     signature dunks.</caption>
    <photographer>Sue Ogrocki</photographer>
    <agency>Reuters</agency>
  </photo>
  <section>
    <par>When it comes to cultivating his public image, Michael Jordan sets the gold standard for
     celebrities . And his latest , and longest, fan dance suggests he hasn't lost a step.</par>
    <par>Everybody, exhale: Tuesday, Jordan officially became willing to once again bring our
     nostalgia to life .</par>
    <par>Jordan, 38, announced he will take his stature off its pedestal and, at least temporarily,
     relinquish his ownership stake and club presidency of the Washington Wizards to again become
     America's most-scrutinized manual laborer.</par>
  </section>
  <section>
    <par>But this time around, it all might look different .</par>
    <par>A new American sensibility, unfolding amid the fight against terrorism, might redefine athletic
     triumphs as something less than truly heroic. Says David Carter, a sports marketer who also teaches at
     the University of Southern California's business school: 'From the top of the food chain to the bottom,
     you're seeing everybody rethink the time, money and energy they give to sports.'</par>
    </section>
</news>
```

Figure 4: Sample news article in XML-based representation.

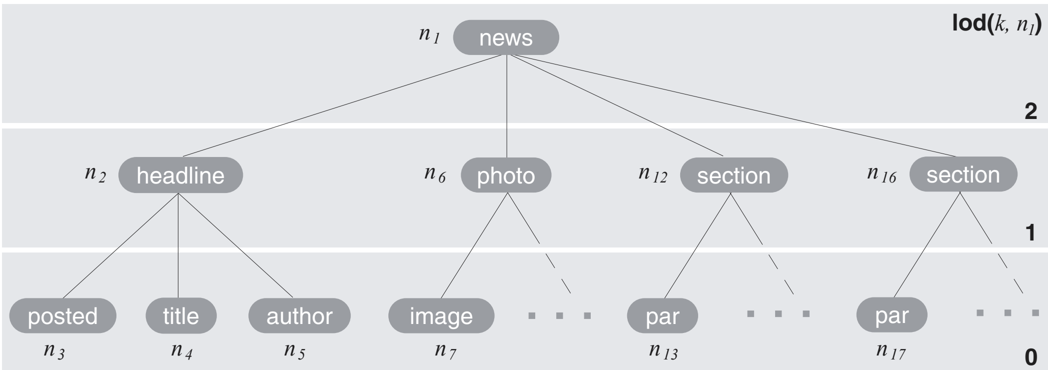


Figure 5: Tree-like document structure of our sample article.

can be ordered as: news > section > par. In practice choosing a level of detail depends on various factors. Typically the interests of a user, the technical constraints of mobile terminals, as well as a usage profile of the e-service are taken into account. Given the limited resources of today's mobile terminals in terms of memory capacity, bandwidth and display capabilities (cf. table 1) the document level for progressive delivery is often chosen low.

2.2 Relevance weighting

Prior to the progressive delivery of a document based on a relevance weighting we determine its natural reading sequence. The reading sequence of a document is defined as the list of nodes at the chosen level of detail as they appear in the author intended sequential reading of the document. In the following let R denote this natural reading sequence of a document. For our sample document at level 1 we have $R = [n_2, n_6, n_{12}, n_{16}]$. The elements of R then have to be reordered for a relevance-based progressive delivery. Thus all document nodes are weighted according to their relevance to the user and subsequently sorted in descending order yielding in the delivery sequence of the document. Finally this sequence is used to deliver the document progressively, i.e. node by node, to the mobile user.

The most important part of relevance weightings are the so-called user interest profiles. We use a standard method based on the well-known vector space model [3] to measure the semantic relevance of textual document parts. In this approach user interests are modeled as shown in table 2. Here, interest vectors express the relative interest of a user in a certain topic, e.g. user Bill is mainly interested in news on the events following September 11 and the "war against terrorism" (90%); while he shows a minor interest for the basketball team "Wizards" (10%), financial news are of no interest for him (0%). Interest profiles are coded in the form of vectors bearing the relative user interest on each topic. In our running example (cf. table 2) the vectors $p_1 = (0.5, 0.5, 0)$, $p_2 = (0.9, 0, 0.1)$ and $p_3 = (0.1, 0, 0.9)$ are associated with our sample users Cathy, Bill and Michael.

#User	Name	911–WAT	FINANCIAL	WIZARDS
1	Cathy	0.5	0.5	0
2	Bill	0.9	0	0.1
3	Michael	0.1	0	0.9

Table 2: Interest profiles of different users.

We assign a relevance weighting to each textual leave of the document that is considered for progressive delivery. This relevance weighting expresses how close the partial document is related to the specific interest profile of a user. Therefore a set of keywords is assigned to each topic with each keyword adorned with a value that reflects the significance of its occurrence. Table 3 displays some of the keywords used to model the three topics 911–WAT, FINANCIAL and WIZARDS in our sample application, e.g. in 911–WAT value 0.5 is assigned to the term "USA" (denoted as "$\text{USA}_{0.5}$"). Since the keyword "USA" appears twice, i.e. in the keyword lists of category 911–WAT and FINANCIAL, an occurrence of "USA" is considered of minor significance compared to "America" which is assigned exclusively to the category 911–WAT.

911–WAT	FINANCIAL	WIZARDS
$\text{USA}_{0.5}$, war_1, $\text{Washington}_{0.5}$, America_1, terror_1, $\text{Europe}_{0.5}$, Afghanistan_1, threat_1, biohazard_1, $\text{fight}_{0.5}$, anthrax_1	NYSE_1, $\text{USA}_{0.5}$, $\text{Europe}_{0.5}$, Dollar_1, stock_1, exchange_1, Euro_1, DOW_1, DAX_1	basketball_1, NBA_1, play-off_1, Michael_1, Jordan_1, dunk_1, standings_1, Wizards_1, $\text{Washington}_{0.5}$, $\text{fight}_{0.5}$, fan_1

Table 3: Keyword lists model different interest topics.

Using the vector space model the relevance weighting of a textual document node can be measured as the scalar product of the user's interest vector p_i and the relevance vector w of the text node. In our running example the relevance vector of the document leaf n_{13} is $w(n_{13}) = (0, 0, 3.0)$, since none of the keywords from the categories 911–WAT or FINANCIAL occurs in node n_{13} whereas three keywords from WIZARDS ("Michael_1", "Jordan_1" and "fan_1") are found; each having a relevance of 1. For e.g. user Michael we would thus get the relevance weighting of node n_{13} as the scalar product of his interest profile $p_3 = (0.1, 0, 0.9)$ and the node vector $w(n_{13}) = (0, 0, 3.0)$ as 2.7.

Figure 6: Progressive delivery for sample user Bill (911–WAT).

Sample progressive delivery

We will now consider the progressive delivery of our sample article n_1 to user Bill based on the relevance weighting introduced above. We choose 1 as the *lod* for this example delivery. This implies that the sub-documents starting with headline, photo, section are considered as the smallest units of the document which cannot be split any further. Here and throughout the paper we are exploring the mobile delivery of content to the Palmpilot PDA [7] using the AvantGo Web browser for mobile devices [1].

Figure 6 depicts the delivery of article n_1 to Bill's mobile terminal. As reflected by his interest profile $p_2 = (0.9, 0, 0.1)$ user Bill is primarily interested in news from the category 911–WAT. While article n_1 is mainly on Basketball and Jordan's comeback to the NBA, it also includes a passage reflecting some public opinion on professional sport in the light of the events of September 11. It is this very passage (sub-document n_{16}) that will be delivered first to user Bill. However, the delivery solely based on relevance weightings reveals some deficiencies: In the progressive delivery of our sample article to user Bill the relevance weighting of a textual section dominates the delivery sequence of the article: sub-document n_{16} – the endmost textual section – even outweighs the headline of the article (screen 2).

2.3 Author annotations

To overcome the problem of outweighing important structural parts we introduce author annotations as meta-markup tags associated with the document type definitions of an application to suspend an unwanted dominance of certain sub-documents in the delivery process. Author annotations influence the standard order of markup tags (cf. section 2.1) by allowing the creator of a document to assign default weighting factors to certain tags enriching the document model for progressive delivery. In the spirit of transcoding hints used in the MPEG-7 format [8] these annotations express that some parts of a document should be ranked over other parts within the delivery. These annotations state basic structural properties (gained from

```
1   <xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">
2
3    <xsd:element name="news" type="News"/>
4
5    <xsd:complexType name="News">
6     <xsd:sequence>
7      <xsd:element name="headline" type="Headline" weight="5"/>
8      <xsd:element name="photo" type="Photo" minOccurs="0" maxOccurs="unbounded"/>
9      <xsd:element name="section" type="Section" maxOccurs="unbounded"/>
10    </xsd:sequence>
11   </xsd:complexType>
12
13   <xsd:complexType name="Headline">
14    <xsd:sequence>
15     <xsd:element name="title" type="xsd:string">
16     <xsd:element name="subtitle" type="xsd:string" minOccurs="0" maxOccurs="1" weight="10">
17     <xsd:element name="author" type="xsd:string">
18     <xsd:element name="posted" type="xsd:date"/>
19    </xsd:sequence>
20   </xsd:complexType>
.
.
.
```

Figure 7: Author annotated XML schema for Internet news articles.

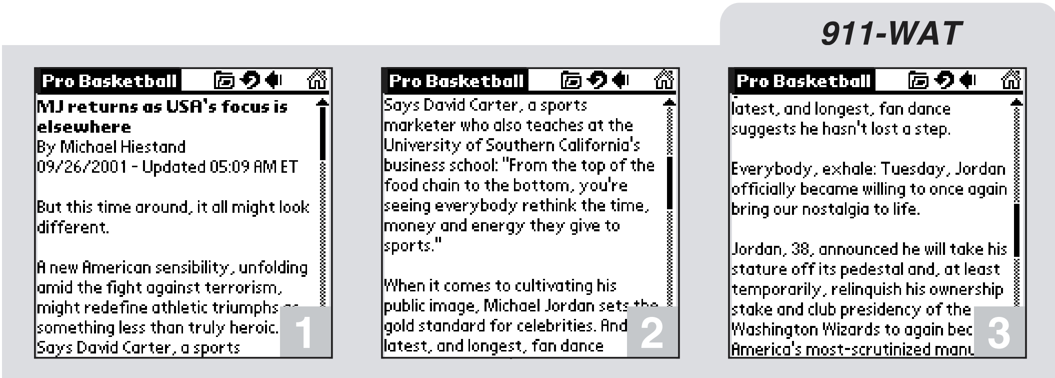
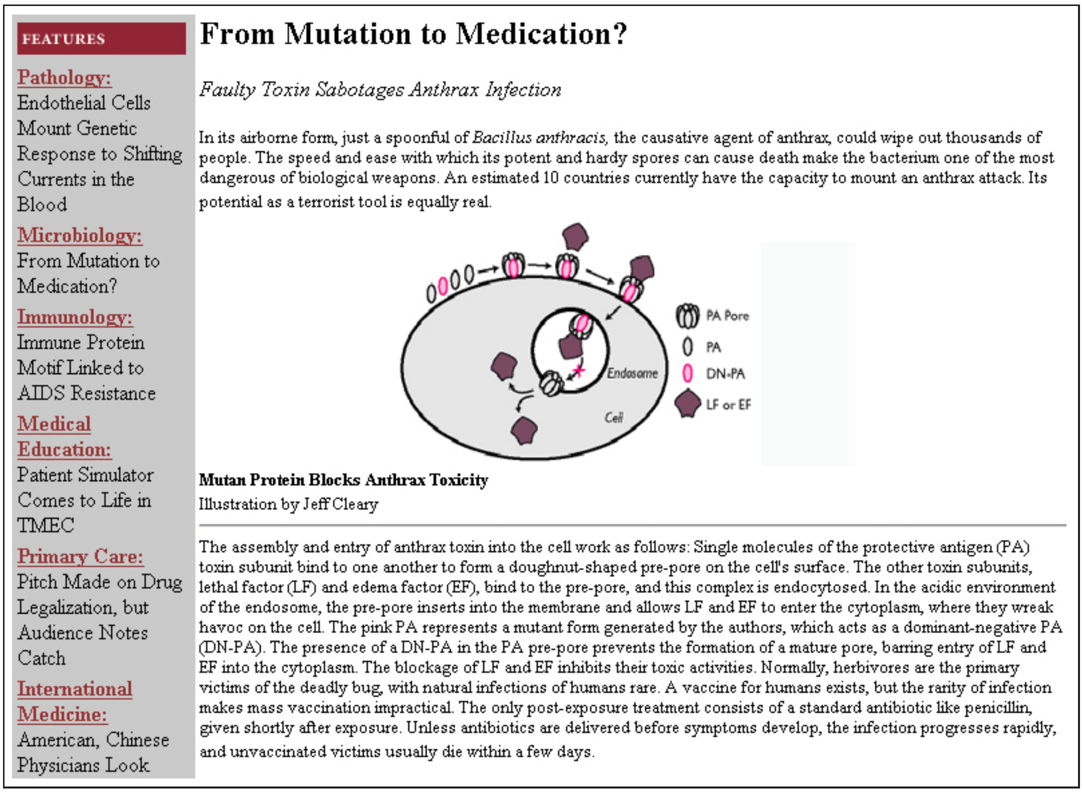


Figure 8: Personalized progressive delivery for sample user Bill.

the document type's semantics) for each document type and will be used in a second step to get a secondary weighting for each block. The two weightings will then be combined to form the total weighting for each block suppressing unwanted dominance of sub-documents in the delivery process.

We illustrate this by an example: figure 7 shows the XML schema used to define articles in our mobile news service. Our schema is enriched by author annotations on the headline and the subtitle tag (line 7 and 16) which assign default weighting factors to these markup tags. These annotations express that by default the headline should always be ranked five times higher as other document parts on the same level of detail, e.g. photo or section. Furthermore, if a headline contains a subtitle, it should be weighted even higher (magnitude of 10). A newly initiated delivery process for user Bill considering the author annotations from above will yield in the delivery depicted in figure 8. The delivered content is personalized according to his interest profile, but this time the headline is delivered first according to the author annotation.

Our approach towards relevance weightings includes textual and structural properties to get an overall satisfying result. However, using the structural semantics via author annotations provides additional benefit: annotations can be used by service providers to support specific usage patterns. Anticipating these patters can essentially improve the quality of service. Consider for instance the restricted Internet capabilities of a modern cell phone. Due to its rather small display, low resolution and low bandwidths it may be preferable to delay or even supress the delivery of images. Thus, if the provider decides that a content



FEATURES

Pathology: Endothelial Cells Mount Genetic Response to Shifting Currents in the Blood

Microbiology: From Mutation to Medication?

Immunology: Immune Protein Motif Linked to AIDS Resistance

Medical Education: Patient Simulator Comes to Life in TMEC

Primary Care: Pitch Made on Drug Legalization, but Audience Notes Catch

International Medicine: American, Chinese Physicians Look

From Mutation to Medication?

Faulty Toxin Sabotages Anthrax Infection

In its airborne form, just a spoonful of *Bacillus anthracis*, the causative agent of anthrax, could wipe out thousands of people. The speed and ease with which its potent and hardy spores can cause death make the bacterium one of the most dangerous of biological weapons. An estimated 10 countries currently have the capacity to mount an anthrax attack. Its potential as a terrorist tool is equally real.

Mutan Protein Blocks Anthrax Toxicity

Illustration by Jeff Cleary

The assembly and entry of anthrax toxin into the cell work as follows: Single molecules of the protective antigen (PA) toxin subunit bind to one another to form a doughnut-shaped pre-pore on the cell's surface. The other toxin subunits, lethal factor (LF) and edema factor (EF), bind to the pre-pore, and this complex is endocytosed. In the acidic environment of the endosome, the pre-pore inserts into the membrane and allows LF and EF to enter the cytoplasm, where they wreak havoc on the cell. The pink PA represents a mutant form generated by the authors, which acts as a dominant-negative PA (DN-PA). The presence of a DN-PA in the PA pre-pore prevents the formation of a mature pore, barring entry of LF and EF into the cytoplasm. The blockage of LF and EF inhibits their toxic activities. Normally, herbivores are the primary victims of the deadly bug, with natural infections of humans rare. A vaccine for humans exists, but the rarity of infection makes mass vaccination impractical. The only post-exposure treatment consists of a standard antibiotic like penicillin, given shortly after exposure. Unless antibiotics are delivered before symptoms develop, the infection progresses rapidly, and unvaccinated victims usually die within a few days.

Figure 9: Web-based presentation of a scientific news article.

is also understandable without images and the quality of service for phones can be increased by generally omitting them, a simple annotations can suppress the delivery of images.

3 Dealing with document fragmentation

Having determined the modified delivery order of a document there are certain cases where the generated delivery sequence is not entirely satisfactory. Especially if a low level of detail is chosen, the partitioning and reordering of the document may cause severe fragmentation, thereby affecting the document's readability. In particular in combination with a poor semantic ranking model which is incapable of capturing the semantic linkage of a textual node to its document context the negative effect of fragmentation can be multiplied.

Let us consider an example which illustrates this behavior: figure 9 depicts a scientific news article from Havard Medical School featuring a strong semantic linkage between the included pictorial illustration ("Mutan Protein Blocks...") and the adjacent text paragraph ("The assembly and entry.."), i.e. the text paragraph mainly provides an explanation of the picture. However, this semantic linkage will neither be expressed by author annotations nor by our textual ranking model. Figure 10 shows the medical article's delivery sequence for our sample user Bill at lod 1 given the XML representation and author annotations from above. Based on his interest profile the document parts related to news category 911–WAT are delivered first. In this case resulting in the separation of the pictorial content placed at the end of the target document (screen 3) and the previously adjacent text paragraph delivered somewhere in the middle (screen

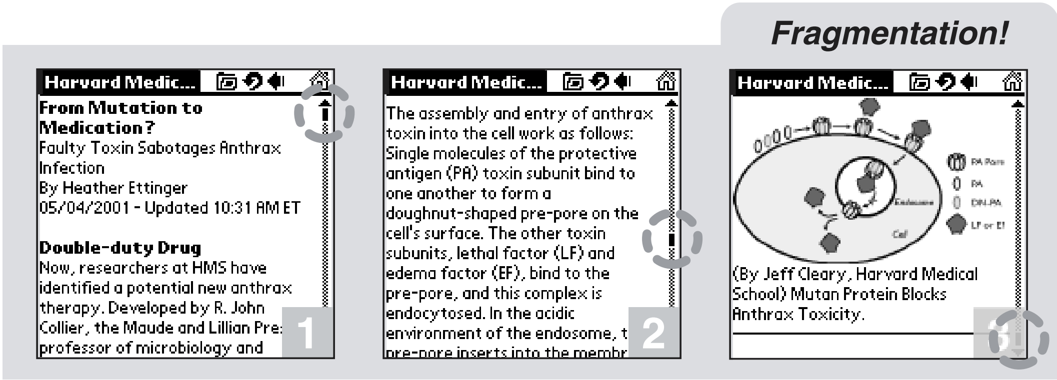


Figure 10: Heavily fragmented delivery for sample user Bill.

2). But since the paragraph mainly consists of a description of the helpful illustration and neither one can be understood without the other, the resulting document's readability is severely affected.

Even though author annotations are useful to express general structural and basic semantic aspects of a document, semantic links between single adjacent parts cannot be modeled adequately to avoid document fragmentation. Thus, in the following we introduce a formal measurement for the degree of fragmentation and in the case of heavy document fragmentation we will use heuristic defragmentation strategies to improve the document's readability prior to its final delivery.

3.1 Measuring fragmentation

To determine the degree of fragmentation of a document, we will introduce the following fragmentation ratios reflecting how the documents readability is affected by the partitioning and reordering of sub-documents. In the following let k be the length of the reading sequence and $\pi : \{1,\ldots,k\} \longrightarrow \{1,\ldots,k\}$ denote the permutation of documents nodes at the chosen level of detail, e.g. with $R = [n_2, n_6, n_{12}, n_{16}]$ and $D = [n_{12}, n_{16}, n_2, n_6]$ as the natural reading and the delivery sequence we have $\pi(1) = 3$ since node n_2 was shifted from the 1st position in R to the 3rd in D.

Fragmentation

The basic ratio of absolute fragmentation F_{abs} is defined as the sum of distances of all document parts in D from their original position in R. This value is normalized to $[0, 1]$ by $\frac{(k-1)^2}{2}$, the upper bound of F_{abs}, to abstract from the document's size (for proof see [11]). Thus we have

$$F_{abs} \quad = \quad \sum_{i=1}^{k} |\pi(i) - i| \qquad \text{and} \qquad F_{rel} \quad = \quad \frac{2}{(k-1)^2} \cdot F_{abs} \, .$$

Incoherence

In addition to measuring how far the absolute positions of nodes are altered by the delivery we define Z_{abs} as the ratio of incoherence. Z_{abs} accounts for distances between nodes in the delivery sequence which were originally adjacent. Again we normalize this value to abstract from the documents size (for proof see

[11]). Thus we have

$$Z_{abs} \quad = \quad \sum_{i=1}^{k-1} |\pi(i+1) - \pi(i)| \qquad \text{and} \qquad Z_{rel} \quad = \quad \frac{2}{k \cdot (k-1)} \cdot (Z_{abs} - (k-1))\ .$$

Readability

Finally we completely abtract from absolute node positions and distances between document nodes. The sequential readability S_{abs} expresses how many nodes of the target document are still readable in their original sequential order. This is determined by iterating over the delivery sequence testing if the original order has been changed. Again we normalize to the relative ratio S_{rel} (for proof see [11]):

$$S_{abs} \quad = \quad \sum_{i=1}^{k-1} \begin{cases} 1 & , \quad \text{if } \pi(i+1) - \pi(i) = 1 \\ 0 & , \quad \text{otherwise.} \end{cases} \qquad , \qquad S_{rel} \quad = \quad \frac{1}{k-1} \cdot S_{abs}$$

3.2 A distance-based strategy for defragmentation

Several defragmentation strategies have been introduced in [11] aiming at the reduction of document fragmentation thereby regarding overall structural characteristics of the target and the source document. These strategies are typically applied directly after the document has been reordered with respect to the relevance weightings and author annotations. Especially for the progressive delivery in mobile environments defragmentation has proven to be of essential importance: due to the restricted capabilities and bandwidth of mobile terminals the level of detail for delivery is often chosen low. However, the lower the level of detail is chosen, the higher is the risk of heavy fragmentation. In the following we will exemplify the idea of defragmentation by the strategy S-DISTANCE from [11].

Defragmentation strategy S-DISTANCE

Figure 11 displays the defragmentation strategy S-DISTANCE. How S-DISTANCE works is best explained by a short example: consider the document structure depicted in figure 12 ready for progressive delivery. At the lowest level of detail we have $R = [3, 4, 5, 7, 8]$ as the natural reading sequence. With relevance weights assigned as indicated in the figure the ad-hoc delivery sequence of the document is $D = [7, 3, 4, 5, 8]$. Note that in D node 7 and 8 (originally adjacent) are separated as far as possible.

The strategy S-DISTANCE uses the distance between nodes adjacent in R with the distance (*dist*) between two nodes being defined as the length of the shortest path between the nodes in the document tree. Figure 12 depicts the first step in the execution of S-DISTANCE illustrating the distances between node 7 (first node in D) and all other nodes in the tree, e.g. $dist(7, 3) = 4$ as the distance between node 7 and 3.

```
strategy S−DISTANCE (D : List(Node))
begin
  first:= head(D); rest:= tail(D);
  foreach (n ∈ rest) do
    w(n):= w(n)/dist(c,first);
  S−DISTANCE (sort(≤, v, rest));
end;
```

Figure 11: S-DISTANCE, focussing on distances between originally adjacent nodes.

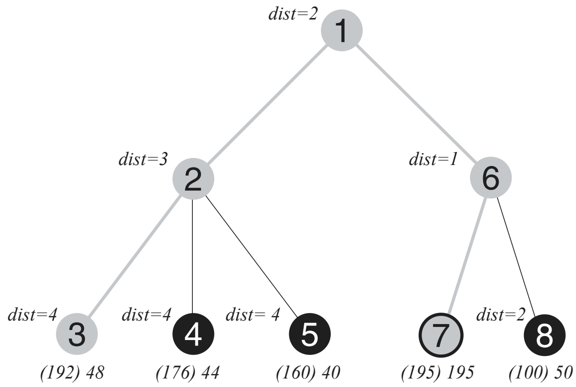


Figure 12: Snapshot of defragmentation using strategy S-DISTANCE.

Using *dist* our strategy iteratively adjusts the weights in D: the *first* node from D is removed (figure 11, line 3). In turn the weights w of all remaining nodes n in the *rest* of D are scaled down by factor $1/dist(n, first)$ (line 4-5). S-DISTANCE is then applied recursively on the *rest* of D which is reordered according to the modified node weights (line 6).

Table 4 shows the complete run of S-DISTANCE for our example from figure 12 and displays the modified weights of the nodes $3, 4, \ldots, 8$ in D together with the modified delivery sequence D' as the strategy's result. Starting with $D = [7, 3, 4, 5, 8]$ the delivery sequence D is recursively processed bringing some nodes together that where originally adjacent, e.g. 7 and 8, and separating some that were originally separated, e.g. 5 and 8. S-DISTANCE terminates with the modified delivery sequence $D' = [7, 8, 3, 4, 5]$ in which parts 7 and 8 – adjacent in the original sequence – again get adjacent.

3.3 Experiments

In addition to small test scenarios as presented in section 2 we have evaluated our framework for progressive delivery in a larger context to prove the effective integration of content selection and defragmentation for real-world applications. We have set up a virtual Internet newspaper containing 15 different news articles. Each article is of roughly the same length and structure as our sports example (cf. figure 3). Using these documents at varying levels of detail 250 different delivery processes for randomly chosen user profiles have been evaluated.

Table 5 shows the average impact of our defragmentation strategy on the fragmentation ratios F_{rel},

INPUT: D	3	4	5	7	8	*OUTPUT: D′*
[7, 3, 4, 5, 8]	192	176	160	**195**	100	[ ]
[8, 3, 4, 5]	48	44	40	**195**	**50**	[7]
[3, 4, 5]	**12**	11	10	**195**	**50**	[7, 8]
[4, 5]	**12**	**5.5**	5	**195**	**50**	[7, 8, 3]
[5]	**12**	**5.5**	**2.5**	**195**	**50**	[7, 8, 3, 4]
[ ]	**12**	**5.5**	**2.5**	**195**	**50**	[7, 8, 3, 4, 5]

Table 4: Complete defragmentation using strategy S-DISTANCE.

STRATEGIES	Ad-hoc	S-DISTANCE	All
F_{abs}	0.68	0.55	0.39
Z_{abs}	0.54	0.09	0.07
S_{abs}	0.02	0.10	0.17

Table 5: Impact of different defragmentation strategies on F_{rel}, Z_{rel} and S_{rel}.

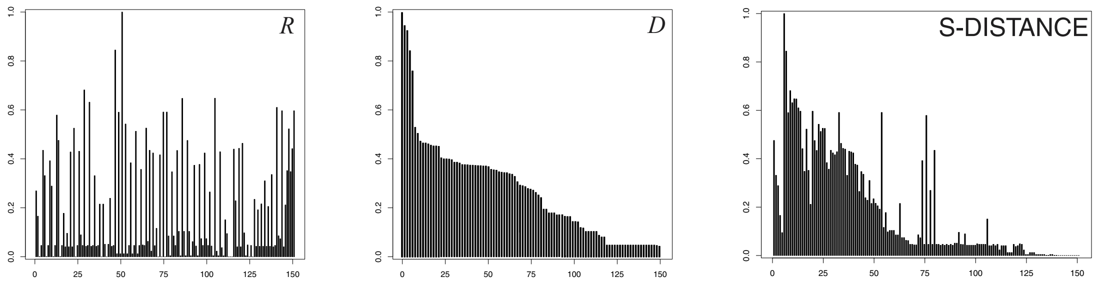


Figure 13: Relevance distribution for original document, ad-hoc delivery and S-DISTANCE.

Z_{rel}, S_{rel} and thus on the readability of the delivered newspaper document. We can observe that the readability is significantly improved on the average, e.g. the document's incoherence measure Z_{abs} is reduced by 83% from 0.54 for ad-hoc delivery (no strategy applied) to 0.09 by the single application of strategy S-DISTANCE; if other defragmentation strategies from [11] are applied Z_{abs} is further reduced to 0.07. Considerable improvements can also be noted for F_{abs} and S_{abs}. Note that for S_{abs} higher values indicate an improved sequential readability.

Relevance weightings under defragmentation

Recall that our goal of content selection has been defined twofold: on the one hand our framework is targeted at the preferred delivery of content-bearing document parts, on the other hand the partitioning and reordering of the document should have only small impact on the document's readability. In the previous section we have already seen that the readability of a document can be significantly improved by defragmentation strategies. But of course we do not want our defragmentation to entirely undo the reordering of our content selection. Thus we will have a look at the relevance weightings assigned by content selection and study the effect of our defragmentation strategies. If our strategies are reasonable, highly weighted document parts should still be delivered preferred in our delivery plan D, but the fragmentation measures should nevertheless be improved.

We will therefore examine the distribution of relevance weightings in the document to be delivered at different stages of content selection. Figure 13 shows the relevance weighting distribution for the Internet newspaper document and a randomly chosen user profile: on the left-hand side the reading sequence R is displayed with relevance weightings somehow distributed over the document parts. The ad-hoc progressive delivery sequence D of the document is shown in the middle of figure 13 since the document was reordered with no defragmentation strategy applied the documents nodes are ordered monotonically according to their relevance weightings. On the right hand side we see the new delivery path after defragmentation.

Without defragmentation the document part with highest relevance weighting is delivered first, then other document parts with monotonic decreasing relevance weightings are delivered. After defragmentation some document parts have been reordered to improve the degree of fragmentation, but we can see that the overall distribution of relevance weightings was not dramatically changed: after improving the document's readability through defragmentation the document parts with high relevance weightings will still be preferred for delivery.

4 Related work

Major tasks of content selection are the personalized adaptation of digital content for specific applications and a progressive delivery of documents under technical constraints. Our approach towards content selection uses XML as a powerful tool for both structure description and semantic annotation. Due to a variety of established standards and tools XML has proven to be well suited for the implementation of value adding Internet services. In [12] a prototypical system for mobile online auctions is presented featuring the synthesis of multimedia documents with suitable file formats for a wide variety of mobile client devices. In this framework all documents are already automatically generated from generic XML sources and even basic usage pattern are implemented.

Most related approaches focus on managing document parts in various file formats with optimized conversions for delivery of entire compound documents. However, all this work focussed on solutions for specific Web browsers. In [9] conversion algorithms for pervasive computing are given. The approach presents media parts of compound documents in form of so-called info-pyramids that allow a systematic view on these objects in different file formats and qualities. Using simple profiles to evaluate different presentations and qualities of the content an optimal multimedia document can be assembled for delivery. However, the document structure in [9] will not be altered: based on technical profiles the approach describes just the replacement of content parts with parts of lower presentation quality.

In contrast [14] presents a semantic evaluation of structured documents for a later reordering of document parts with respect to a certain level of detail. This work focuses on the evaluation of documents using information retrieval techniques. However, the evaluation is not performed with respect to individual user profiles, but relies on word distributions within the document to find its most relevant parts. The concept of [14] is extended in [6] by a protocol for successive and secure transfer of relevant document parts, however without considering technical constraints of client devices or document fragmentation.

5 Summary and outlook

We presented a coherent approach for delivery of multimedia documents to mobile users, its three major steps being the synthesis of documents, the management of the media objects and delivery to the client devices using protocols like WAP or i-mode. In particular we investigated the progressive delivery of digital content in mobile applications, where the content-bearing document parts have to be determined for preferred delivery. Thus users are enabled to get fast access to most relevant information within a document and can decide if the entire document is useful, or the delivery could already be terminated at an early stage. Especially because of low bandwidths and service costs often charged w.r.t. online time the idea of progressive delivery promises an essential improvement for mobile services.

To support progressive delivery we have developed a framework called content selection that takes the user's personal notion of relevant document parts, the document's structure, semantic annotations of the author and the client device's technical profile into account. For details on the consideration of technical constraints see [11]. We exemplified all steps setting up a prototypical mobile Internet news service. With suitable stylesheets and XSLT transformations in our framework generic XML documents can automatically be transformed and a new reordered document is synthesized suitable for delivery. However, practical tests of the above concept in the area of mobile Web have shown that when reordering document parts, the content's readability often is severely affected and if low levels of detail are chosen, documents may even become completely incomprehensible. Thus we also proposed a strategy for defragmentation to improve the documents' readability.

Content selection as presented here relies on user preferences, technical constraints and fundamental structural characteristics of documents. For a proof of concept we have used standard IR methods to model personal user interests and to rank document nodes for progressive delivery. In forthcoming work key components of the system can be enhanced towards a more flexible and intuitive use of preference modelling, including an advanced user profiling, the determination of personalized document delivery paths and the concept of author annotations. Within the research initiative Preference World at the University of Augsburg ways towards intuitive personalization are investigated. The core modeling technique is a universal strict partial order semantics for preferences which closely matches people's intuition [4]. Various portions of this presented preference model have already been prototyped or are in commercial use in SQL or XML environments.

References

[1] Inc. AvantGo. Mobile Internet Service. http://www.avantgo.com, 2001.

[2] World Wide Web Consortium. Extensible Markup Language (XML) 1.0. http://www.w3.org/.

[3] F. Crestani, M. Lalmas, and C.J. van Rijsbergen, editors. *Uncertainty and Logics: Advanced Models for the Representation and Retrieval of Information*. International Series on Information Retrieval. Kluwer Academic, September 1998.

[4] W. Kießling. Foundations of Preferences in Database Systems. Technical Report 2001-08, Universität Augsburg, Institut für Informatik, October 2001.

[5] W. Kießling, W.-T. Balke, and M. Wagner. Personalized Content Syndication in a Preference World. In *Proceedings of the EnCKompass Workshop on E-Content Management (ENCKOMPASS)*, Eindhoven, The Netherlands, June 2001.

[6] H. V. Leong, D. McLeod, A. Si, and S. Yau. On Supporting Weakly-Connected Browsing in a Mobile Web Environment. In *Proceedings of the 20th International Conference on Distributed Computing Systems, (ICDCS)*, pages 538–546, Taipei, Taiwan, 2000.

[7] Inc. Palm. Palmpilot Personal Digital Assistant. http://www.palm.com, 2001.

[8] J. R. Smith. MPEG-7 Standard for Multimedia Databases. In Walid G. Aref, editor, *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, page 627, Santa Barbara, CA, USA, May 2001. ACM Press.

[9] J. R. Smith, R. Mohan, and C.-S. Li. Scalable Multimedia Delivery for Pervasive Computing. In *Proceedings of the ACM International Conference on Multimedia*, pages 131–140, Orlando, FL, USA, 1999.

[10] USA Today. News & Information Homepage. http://www.usatoday.com/, 2001.

[11] M. Wagner. *Multimedia Delivery für personalisierbare E-Services*. PhD thesis, University of Augsburg, Institute of Computer Science, 2002.

[12] M. Wagner, W.-T. Balke, and W. Kießling. An XML-based Multimedia Middleware for Mobile Online Auctions. In *Proceedings of the 3rd International Conference on Enterprise Information Systems (ICEIS)*, pages 934–944, Setúbal, Portugal, July 2001.

[13] M. Wagner, S. Holland, and W. Kießling. Efficient and Flexible Web Access to Art-Historical Image Collections. In *Proceedings of the 15th ACM Symposium on Applied Computing (SAC)*, volume 2, pages 915–921, Como, Italy, 2000.

[14] S. Yau, H. V. Leong, D. McLeod, and A. Si. On Multi-Resolution Document Transmission in Mobile Web. *SIGMOD Record*, 28(3):37–42, 1999.