On Multi-resolution Document Transmission in Mobile Web

Stanley M.T. Yau† Hong Va Leong† Dennis McLeod‡ Antonio Si⋆
†Department of Computing, Hong Kong Polytechnic University, Hung Hom, Hong Kong
‡Computer Science Department, University of Southern California, Los Angeles, CA 90089
⋆Sun Microsystems, 901 San Antonio Road, Palo Alto, CA 94303

Abstract

We propose a multi-resolution transmission mechanism *that allows various organizational units of a web document to be transferred and browsed according to the amount of information captured. We define the notion of* information content *for each individual organizational unit of a web document as an indication of its captured information. The concept of information content is used as a foundation for defining the notion of* relative information content *which determines the transmission order of various units. Our mechanism allows a web client to explore the more content-bearing portion of a web document earlier so as to be able to terminate browsing a possibly irrelevant document sooner. This scheme is based on our observation that different organizational units of a document contribute to different amount of information to the document. Such a multi-resolution transmission paradigm is particularly useful in mobile web where the wireless bandwidth is a scarce resource and browsing every document in detail would consume the bandwidth unnecessarily. This is becoming more serious when the size of a web document is getting large, such as technical documents. We then present a prototype of the system in Java and CORBA to illustrate its feasibility.*

1 Introduction

We focus on a mobile environment in which mobile clients navigate inter-related web documents via common browsers. We term such an environment, a *mobile web environment*. Since communication between mobile clients and web server is via low bandwidth wireless channels, traffic generated due to web accesses should consume as little bandwidth as possible. In this aspect, conventional approaches to web navigation suffer from serious limitations.

Conventional approaches to web navigation usually involve the searching of web documents via some search engines, followed by human exploration of each document for relevance. By a document, it not only refers to a single web page, but it may also include a collection of hierarchically linked related pages, composing a larger document. We observe quite often that most documents identified by a search engine are irrelevant to a user; this not only wastes precious bandwidth for transferring the documents, but also consumes the limited energy of a mobile client unnecessarily. This problem becomes more serious when the size of a web document is getting bigger as we witness a proliferation of technical documents published to the web.

We address the issue of mobile web browsing through a *multi-resolution transmission* paradigm in [9]. It is built upon the XML markup language [6] which defines a structure for web documents. It is easily extensible to cater for HTML documents. The multi-resolution scheme is based on the observation that different organizational units of a document contribute to different amount of information content. According to the XML structure of a document, the scheme partitions a document into multiple organizational units, defining various *Levels Of Detail* (*LOD*). A notion of *information content* is associated with each unit, indicating its amount of information captured. Units of higher information content will be transmitted earlier. This technique allows a document to be transmitted and browsed at a coarser resolution, with the details to be filled in progressively. A user could thus decide early if the document is of any interest. This could reduce scarce wireless bandwidth consumption by terminating early the transmission of irrelevant documents. In this paper, we extend the technique to allow a mobile client to perceive the maximum amount of information content of a document within the shortest possible duration. This is achieved by the notion of *relative information content* for defining the transmission order. The feasibility of our approach is demonstrated with a prototype.

The remainder of this paper is organized as follows. Section 2 is devoted to a brief survey of previous work on mobile information access and document browsing. In Section 3, we present our multi-resolution scheme in detail. The implementation of our prototype is described in Section 4. Finally, we offer brief concluding remark and future research directions in Section 5.

2 Related Work

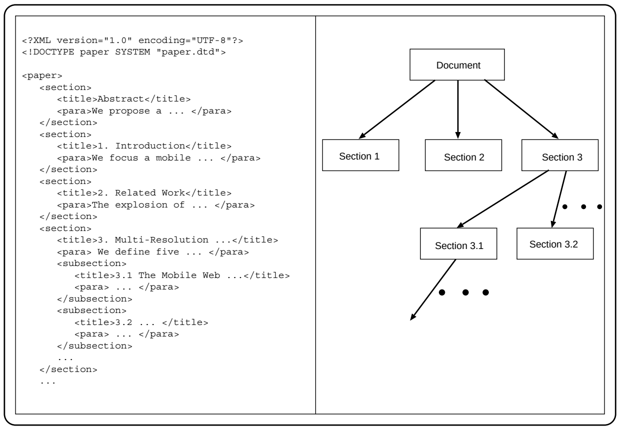
The explosion of information available on the Internet and the user-friendliness of web browsers have dramatically changed the way information is accessed and delivered. New web users will simultaneously experience the excitement of boundless information and the frustration of trying to find what they actually want. Much work has focused on the WYGIWYW (**W**hat **Y**ou **G**et **I**s **W**hat **Y**ou **W**ant) paradigm. A common technique to improve information searching is to build an index over a collection of documents found by a web search process such as Lycos [11] or WebCrawler [13], which typically searches exhaustively rather than to fulfill a particular query. A probably better approach is to establish a user profile, capturing individual users' interests. The profile is used to filter out irrelevant information identified by a search engine [1, 5]. Mechanisms for updating the profile are also provided so that it could be adapted to changes in user interest. Rather than providing a user with selected documents, the WebWatcher system [3] assists a user in his/her browsing behavior. It interactively offers advice about which subsequent hyperlink(s) would likely contain the most relevant information.

Advances in wireless communication and portable computers have enabled users to access web information along the road [12]. Since wireless channels have limited bandwidth and mobile clients are constrained by limited battery life, the challenge is to utilize the scarce bandwidth and power carefully, striking a balance for the best solution for the application at hand [2, 8]. To reduce energy consumption, clock rate reduction and disk spin-down techniques have been proposed [15]. To reduce bandwidth utilization, techniques for caching of data items from the server in a client's local storage have been investigated [7, 10], in reducing wireless traffic and the dependency on the wireless channels.

3 Multi-Resolution Transmission

We define five LOD commonly found in most documents: document, section, subsection, subsubsection, and paragraph, though an arbitrary number of LOD could be defined. These LOD follow a total order: paragraph $\lhd$ subsubsection $\lhd$ subsection $\lhd$ section $\lhd$ document. The document level constitutes the highest degree of detail, containing the largest amount of information while the paragraph level constitutes the lowest degree of detail, capturing the least amount of information. For instance, given a paragraph from a book (collections of chapters or documents), one might not be able to tell much about the content of the chapter (document). A section as modeled by the section LOD may be declared within a pair of <section> and </section> tags in an XML document of type, say, ResearchPaper, defined in a DTD (Document Type Definition) file. Other organizational units may be modeled similarly. Our definition of LOD is also applicable to HTML documents, by defining a mapping between a LOD to a corresponding HTML tag. For instance, for an HTML document, a section LOD might be defined with a <H1> tag.

Using a higher LOD might be more intuitive but with a lower LOD, there would be more flexibility for the smaller organizational units to be rearranged for transmission. In practice, transmitting and browsing at section or subsection LOD will be more beneficial, allowing a mobile client to grasp a reasonable detail of content from the document, while at the same time, be able to terminate the transmission of an irrelevant document early.



```
<?XML version="1.0" encoding="UTF-8"?>
<!DOCTYPE paper SYSTEM "paper.dtd">

<paper>
   <section>
      <title>Abstract</title>
      <para>We propose a ... </para>
   </section>
   <section>
      <title>1. Introduction</title>
      <para>We focus a mobile ... </para>
   </section>
   <section>
      <title>2. Related Work</title>
      <para>The explosion of ... </para>
   </section>
   <section>
      <title>3. Multi-Resolution ...</title>
      <para> We define five ... </para>
      <subsection>
         <title>3.1 The Mobile Web ...</title>
         <para> ... </para>
      </subsection>
      <subsection>
         <title>3.2 ... </title>
         <para> ... </para>
      </subsection>
      ...
   </section>
   ...
```

Figure 1. Sample document and document tree

Each web document is described by a semantic structure called the *structural characteristic* (SC), a tree-like structure modeling the structural organization of a document. It defines the transmission order of the document at the required LOD. Figure 1 illustrates a sample document and its corresponding document tree, defining the SC. As shown in the figure, the root node of the SC models the document itself. The children of the root node model the sections of the document. Similarly, the children of a section node model the subsections. Each node, n_i, is associated with an information content, p_i, which indicates the amount of information content captured within the organizational unit it models.

To transmit a document at a required LOD, the nodes at that LOD are ordered according to their information contents. This order defines the transmission order of the corresponding organizational units of the document. The higher the information content of an organizational unit, the earlier the unit is transmitted. The idea is to allow users to perceive the content of the document as early as possible. When a unit is transmitted, all its sub-units will be transmitted in a sequential manner. Notice that transmitting a document at the document LOD in our algorithm is equivalent to the conventional transmission and browsing paradigm as all sections and paragraphs are transmitted sequentially.

Transmitting a document at the lowest LOD at the paragraph level allows a user to have a very quick glance at the rough content of the document. By contrast, if the highest LOD about the document is preferred, the document should be transmitted at the document level.

3.1 The Mobile Web Architecture

The high level mobile web architecture [9] is depicted in Figure 2. A mobile client initiates a request for a web document, say, an XML document, via a web browser, which interacts with a web server via HyperText Transfer Protocol. The web server will contact a database gateway process, composed of two modules: document transmitter and structural characteristic generator. The document transmitter looks up the database to locate the SC of the requested document. If the SC exists in the database, the document transmitter retrieves it from the database and transfers the requested document at the specified LOD as defined by the retrieved SC. Otherwise, the transmitter will inform the structural characteristic generator to generate the SC of the web document and store it in the database subsequently.

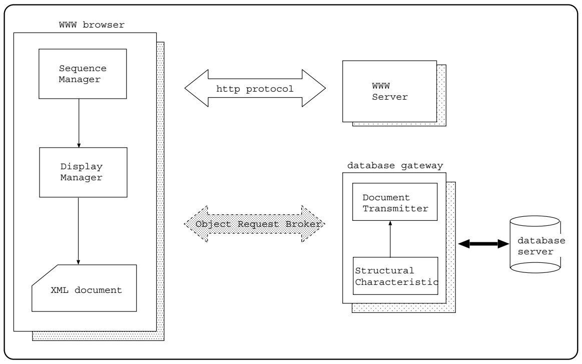


Figure 2. Architecture of our system

A mobile client contains two modules: sequence manager and rendering manager. The sequence manager is responsible for receiving an organizational unit of the requested document from the database gateway. It will determine the position with respect to the document where the received unit should be rendered. The positioning information and the unit received are passed to the rendering manager, which renders the unit in a web browser at the appropriate position.

3.2 Structural Characteristic Generation

To establish an SC for a document, the document is preprocessed and a keyword-based logical index is established for each organizational unit. The SC is created by deriving the information content of each organizational unit from the logical index. The pre-processing process is organized into five modules: *document recognizer*, *canonical converter*, *word filter*, *keyword extractor*, and *structural characteristic generator*, operating in a pipelined fashion.

The document recognizer converts an XML document into a plain ASCII document, utilizing the DTD definition for the structure of the document. The canonical converter then converts document words into their canonical form [14]. The word filter eliminates non-meaning-bearing close-class words. The keyword extractor performs a frequency analysis on the potential keywords. Words with a frequency falling within the *threshold interval* will be extracted as keywords. In addition, certain specially formatted words, such as boldfaced and italized, also qualify as keywords. By adjusting the threshold interval, the extraction process is flexible enough to adapt to different collections of documents. The structural characteristic generator computes the information content of each organizational unit and generates the SC.

3.3 Information Content

The information content, p_i, of an organizational unit n_i is measured by the amount of information captured within each keyword of the unit. Keywords which occur more frequently than others should carry more information of the document, but each occurrence of those high frequency keywords will carry less information. Each keyword, a, is associated with a weight, ω_a. Each particular occurrence of a may be in different context and thus, warrant different contribution. For instance, an occurrence inside a section heading (<section>) or inside an italized format (<I>) should be considered more important. We associate a *context adjustment score* with each occurrence of a, defaulted to 1. For example, the context adjustment score for the italized format may be 3. Such a contextual adjustment can be generalized for co-occurrence of keywords, as a co-occurrence matrix $\mathcal{C}$. For instance, if "Hong" is related to "Kong" as in the phrase "Hong Kong", we can set the co-occurrence adjustment of the two words to a value larger than 1. Each "Hong" or "Kong" alone will be treated as normal keywords, but if they occur together, the information content is much higher. For negatively correlated words, the co-occurrence adjustment would be less than 1. The information content, p_i, of n_i is the sum of the weights of words multiplied by the context adjustment score in the unit.

For notational convenience, we denote the number of occurrences of a keyword a, after contextual adjustment, in a document, D, by $|a_D|$ and the number of occurrences of a after contextual adjustment in an organizational unit, n_i, by $|a_{n_i}|$. The occurrence vector, V_D, of the set of keywords in D, $A_D = \{a | a \text{ is a keyword in } D\}$, can be represented as $V_D = \{|a_D| \mid a \in A_D\} = \{v_1, v_2, ..., v_{|A_D|}\}$. We use a logarithmic function to define the weight, ω_a, of each occurrence of a keyword, a, in D. This function could model the decay in importance of a with respect to its increase in frequency in D. It is defined as $\omega_a = 1 - \log_2(|a_D|/||V_D||)$ where $||V_D||$ is the norm of the occurrence vector V_D. We adopt the infinity norm $||V_D||_\infty = max(v_i)$ here.

Definition 1 *The information content* p_i *of* n_i *is*

$$p_i = \frac{\sum_{\forall \text{keyword } a \in n_i} |a_{n_i}| \omega_a}{\sum_{\forall \text{keyword } d \in D} |d_D| \omega_d}.$$

A property of information content is that it follows the additive rule, i.e., for a given node, n_i, with children, $\{n_{i,1}, \cdots, n_{i,m}\}$, $p_i = \sum_{k=1}^{m} p_{i,k}$. It is not difficult to see why the additive rule holds from the definition.

There are also other ways to determine keywords other than frequency analysis, such as co-occurrence frequency of multiple words within a phrase. We just describe one possible approach for simplicity here. Nonetheless, the vector space model has been shown to be competitive with alternative methods [4].

3.4 Total and Relative Information Content

The notion of information content enables a client to perceive the largest amount of information at the required LOD. Since a mobile environment is of limited bandwidth and is vulnerable from disconnection, a client needs to obtain the required information within the shortest possible period of time. Given a specific LOD, we define the notion of *relative information content* to determine the optimal transmission order for organizational units at the required LOD such that a client could perceive more information within the shortest possible duration. We start by defining the *perceived information content* at a particular time, t.

Definition 2 *The perceived information content observed at a client at time t is the amount of information it can read, i.e.,* $P(t) = \sum_{n_i \text{ in client}} p_i$.

Definition 3 *The total information content perceived by a client over a period of time* $[t_1, t_2]$ *is the total amount of perceived information content available to the client over the period, i.e.,* $\tilde{P} = \int_{t_1}^{t_2} P(t)dt$.

Definition 4 *The expected information content over a period* $[t_1, t_2]$ *is the total information content normalized over the period of time, i.e.,* $\overline{P} = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} P(t)dt$.

Let the time taken to transmit a unit n_i of length l_i be L_i. We assume a *uniform* network transmission delay and a *uniform* distribution of weights within a unit. If n_i is the first unit to be transmitted, the total information content perceived by a client by browsing at n_i would be $\frac{1}{2}p_i L_i$, with an expected content of $\frac{1}{2}p_i$.

Consider the browsing session for document D consisting of m organizational units, $D = \{n_1, n_2, ..., n_m\}$. We represent a transmission order of the units by a permutation, $\pi = \langle j_1, j_2, ..., j_m \rangle$, denoting the transmission order of $\langle n_{j_1}, n_{j_2}, ..., n_{j_m} \rangle$. The total information content when D is transmitted over the period is

$$\begin{aligned}
\tilde{P}_D &= \frac{1}{2}p_{j_1}L_{j_1} + (p_{j_1}L_{j_2} + \frac{1}{2}p_{j_2}L_{j_2}) + \\
&\quad (p_{j_1}L_{j_3} + p_{j_2}L_{j_3} + \frac{1}{2}p_{j_3}L_{j_3}) + ... + \\
&\quad (p_{j_1}L_{j_m} + p_{j_2}L_{j_m} + ... + p_{j_{m-1}}L_{j_m} + \frac{1}{2}p_{j_m}L_{j_m}) \\
&= \sum_{i=1}^{m}(p_{j_i}\sum_{k=i}^{m}L_{j_k}) - \frac{1}{2}\sum_{i=1}^{m}p_{j_i}L_{j_i}.
\end{aligned}$$

The expected information content $\overline{P}_D$ is thus $\tilde{P}_D / \sum_{i=1}^{m} L_i$. Different transmission orders would yield different values of $\overline{P}_D$. Our objective is not only to maximize the information content transmitted, but also the total and expected information content transmitted. We, thus, have to determine an optimal transmission order, π_{opt}, for transmitting D at the required LOD. To achieve this goal, we define the notion of *relative* information content of an organizational unit.

Definition 5 *The relative information content,* γ_i, *of organizational unit* n_i *is defined as the ratio of its information content to its length, i.e.,* $\gamma_i = p_i / l_i$.

Theorem 1 *Given a set of organizational units* $D = \{n_1, n_2, ..., n_m\}$, *with relative information contents* $\gamma_1, \gamma_2, ..., \gamma_m$ *respectively, the transmission order* $\pi = \langle j_1, j_2, ..., j_m \rangle$ *such that* $\gamma_{j_1} \geq \gamma_{j_2} \geq ... \geq \gamma_{j_m}$ *will maximize the expected information content at a client.*

Proof: Since L_i is proportional to l_i ($L_i = cl_i$ where c is a proportional constant), we express the expected information content in terms of γ_i as

$$\overline{P} = \frac{1}{\sum_{i=1}^{m} c/\gamma_i}[\sum_{i=1}^{m}(p_{j_i}\sum_{k=i}^{m}\frac{cp_{j_k}}{\gamma_{j_k}}) - \frac{1}{2}\sum_{i=1}^{m}\frac{cp_{j_i}^2}{\gamma_{j_i}}].$$

We observe that the second summation term is constant regardless of the permutation, so is the divisor. Removing the constant terms, this is equivalent to maximizing

$$\sum_{i=1}^{m}(p_{j_i}\sum_{k=i}^{m}\frac{p_{j_k}}{\gamma_{j_k}}).$$

With $D = \{n_1, n_2\}$ as base case, it is easy to verify that sending n_1 before n_2 leads to higher expected information content if $\gamma_1 > \gamma_2$. The reversed order is better if $\gamma_2 > \gamma_1$. The expected information content of the two transmission orders are equal if $\gamma_1 = \gamma_2$. With the induction hypothesis for the number of units $m = k$ in D, when $m = k + 1$, there are only $k + 1$ possible optimal candidates to consider out of the $(k+1)!$ permutations, with the first k elements in the permutation sorted in γ. We can then show that Theorem 1 holds. □

3.5 Sample Structural Characteristic

We demonstrate the SC generation with a draft of this manuscript. The information contents and relative information contents of the organizational units of this document are illustrated in Table 1. Keywords within section headings and <B> and <I> tags receive a context adjustment score of 3. The abstract is considered as Section 0. For consistency and clarity purpose, paragraphs not belonging to any subsection are grouped under a virtual subsection. For instance, all paragraphs

Sect.	/ Subsect.	/ Para.	Content p	Relative Cont. γ
0			0.05172	0.47981×10^{-4}
	0.0		0.05172	0.47981×10^{-4}
		0.0.0	0.05172	0.47981×10^{-4}
1			0.12488	0.50192×10^{-4}
	1.0		0.12488	0.50192×10^{-4}
		···	···	···
2			0.09521	0.41395×10^{-4}
	2.0		0.09521	0.41395×10^{-4}
		···	···	···
3			0.55939	0.44591×10^{-4}
	3.0		0.12369	0.41901×10^{-4}
		3.0.1	0.04080	0.49332×10^{-4}
		3.0.2	0.01205	0.27883×10^{-4}
		3.0.3	0.03619	0.52223×10^{-4}
		3.0.4	0.03466	0.34658×10^{-4}
	3.1		0.15849	0.56144×10^{-4}
		3.1.1	0.04455	0.56325×10^{-4}
		3.1.2	0.02220	0.48051×10^{-4}
		3.1.3	0.03740	0.86563×10^{-4}
		3.1.4	0.05435	0.47756×10^{-4}
	3.2		0.09863	0.43336×10^{-4}
		3.2.1	0.06881	0.45238×10^{-4}
		3.2.2	0.02982	0.39503×10^{-4}
	3.3		0.11849	0.40208×10^{-4}
		···	···	···
	3.4		0.06008	0.38835×10^{-4}
		···	···	···
4			0.10704	0.47094×10^{-4}
	4.0		0.10704	0.47094×10^{-4}
		···	···	···
5			0.06176	0.37566×10^{-4}
	5.0		0.06176	0.37566×10^{-4}
		···	···	···

Table 1. Information content of this manuscript

belonging to Section 3, but not to Subsection 3.1 are grouped under the virtual Subsection 3.0.

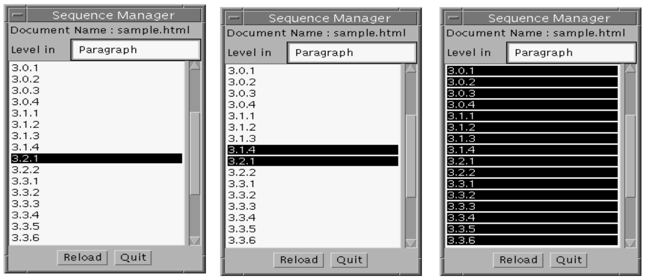
The transmission order of the document at section LOD based only on information content is ⟨Section 3, Section 1, Section 4, Section 2, Section 5, Section 0⟩. Based on relative information content, the order would be ⟨Section 1, Section 0, Section 4, Section 3, Section 2, Section 5⟩. Thus, Introduction and Abstract are detected to be more important and transmitted earlier by using relative information content to determine the transmission order. If subsection LOD is preferred, some subsections of Section 3 still contain more information. The transmission is at the order of ⟨Subsection 3.1, Subsection 1.0, Subsection 3.0, Subsection 3.3, ···⟩ if only information content is used. With relative information content, the order is ⟨Subsection 3.1, Subsection 1.0, Subsection 0.0, Subsection 4.0, ···⟩.

4 Implementation and Prototype

Our prototype is implemented as a CORBA architecture. The database gateway process in Figure 2 is implemented as a CORBA server, using Java. We employ Oracle database system to maintain the SCs of each document. The communication between Oracle and the database gateway is via JDBC protocol while the communication between the gateway process and a mobile client is via Object Request Broker. Implementing the database gateway as a CORBA server could lead to better performance than as a CGI process. Employing JDBC for communicating with the database server provides interoperability with other databases.

A mobile client must render a document based on the order of the progressive transmission, which depends on the requested LOD. Our interface for rendering is implemented as a Java applet, allowing our client to be run on various web browsers and platforms. Our mobile web client interface takes the look as in Figure 4.

We illustrate a sample rendering session using this manuscript, based on the information content of the organizational units. The user interface for transmission based on relative information content is similar, though the order could be different. As described in Section 3.5, the order when transmitted at the section LOD is ⟨Section 3, Section 1, Section 4, Section 2, Section 5⟩. The order at the paragraph LOD would be ⟨Paragraph 3.2.1, Paragraph 3.1.4, Paragraph 0.0.0, ···⟩. The content displayed at two different stages after Paragraph 3.2.1 and Paragraph 3.1.4 of Section 3 have been received and rendered is depicted in Figures 4a and 4b. To give users a better idea of what is being transmitted and what has been transmitted, the sequence manager shows the organizational units that have been received and rendered, as illustrated in Figure 3. Figures 3a and 3b correspond to the states of the browser in Figures 4a and 4b respectively. Figure 3c represents the final state when all paragraphs have been rendered.



Document Name : sample.html
Level in: Paragraph

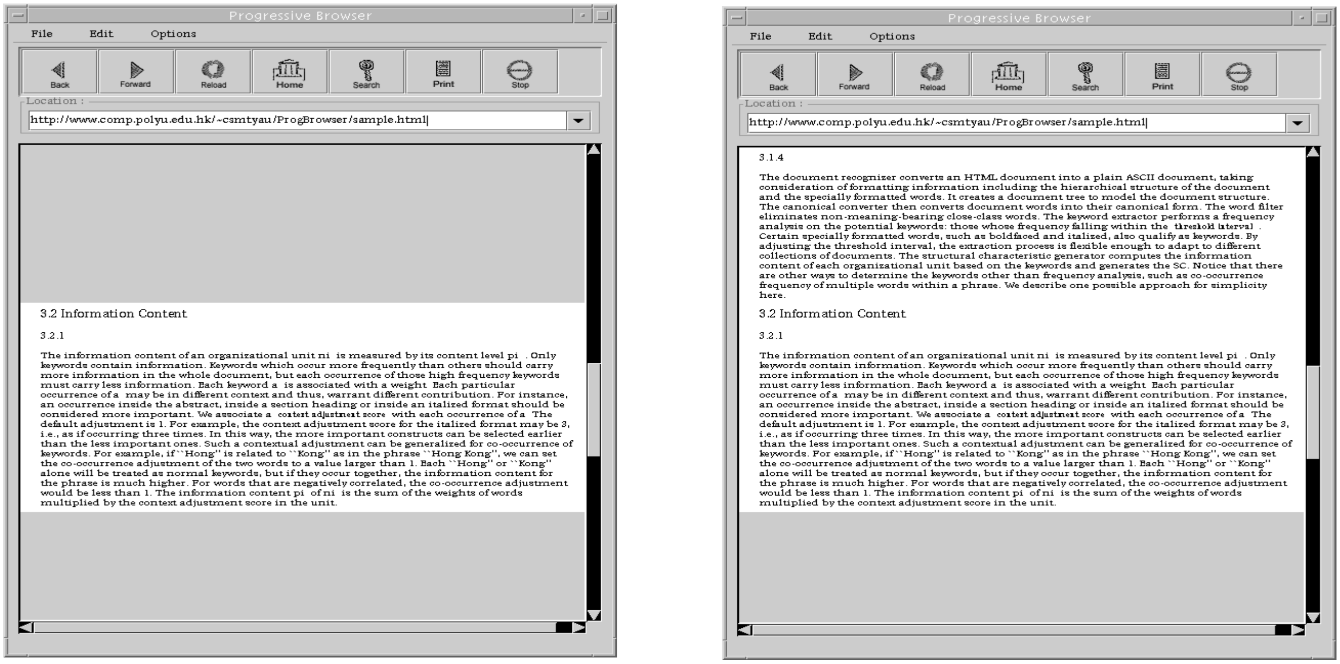
(a) (b) (c)

Figure 3. Status panel during browsing

5 Conclusion

We have presented the design and implementation of a framework for transmitting and browsing web documents at various LOD. Based on information content, it presents users with the main document content before presenting supplementary information. Relative information content is introduced to identify the optimal transmission order, maximizing the perceived information content. This is especially important in a mobile environment where bandwidth is scarce.

Currently, our prototype requires a well-defined organizational structure on a web document. However, in



http://www.comp.polyu.edu.hk/~csmtyau/ProgBrowser/sample.html

3.2 Information Content

3.2.1

The information content of an organizational unit ni is measured by its content level pi . Only keywords contain information. Keywords which occur more frequently than others should carry more information in the whole document, but each occurrence of those high frequency keywords must carry less information. Each keyword a is associated with a weight Each particular occurrence of a may be in different context and thus, warrant different contribution. For instance, an occurrence inside the abstract, inside a section heading or inside an italized format should be considered more important. We associate a context adjustment score with each occurrence of a The default adjustment is 1. For example, the context adjustment score for the italized format may be 3, i.e., as if occurring three times. In this way, the more important constructs can be selected earlier than the less important ones. Such a contextual adjustment can be generalized for co-occurrence of keywords. For example, if ``Hong'' is related to ``Kong'' as in the phrase ``Hong Kong'', we can set the co-occurrence adjustment of the two words to a value larger than 1. Each ``Hong'' or ``Kong'' alone will be treated as normal keywords, but if they occur together, the information content for the phrase is much higher. For words that are negatively correlated, the co-occurrence adjustment would be less than 1. The information content pi of ni is the sum of the weights of words multiplied by the context adjustment score in the unit.

http://www.comp.polyu.edu.hk/~csmtyau/ProgBrowser/sample.html

3.1.4

The document recognizer converts an HTML document into a plain ASCII document, taking consideration of formatting information including the hierarchical structure of the document and the specially formatted words. It creates a document tree to model the document structure. The canonical converter then converts document words into their canonical form. The word filter eliminates non-meaning-bearing close-class words. The keyword extractor performs a frequency analysis on the potential keywords: those whose frequency falling within the threshold interval . Certain specially formatted words, such as boldfaced and italized, also qualify as keywords. By adjusting the threshold interval, the extraction process is flexible enough to adapt to different collections of documents. The structural characteristic generator computes the information content of each organizational unit based on the keywords and generates the SC. Notice that there are other ways to determine the keywords other than frequency analysis, such as co-occurrence frequency of multiple words within a phrase. We describe one possible approach for simplicity here.

3.2 Information Content

3.2.1

The information content of an organizational unit ni is measured by its content level pi . Only keywords contain information. Keywords which occur more frequently than others should carry more information in the whole document, but each occurrence of those high frequency keywords must carry less information. Each keyword a is associated with a weight Each particular occurrence of a may be in different context and thus, warrant different contribution. For instance, an occurrence inside the abstract, inside a section heading or inside an italized format should be considered more important. We associate a context adjustment score with each occurrence of a The default adjustment is 1. For example, the context adjustment score for the italized format may be 3, i.e., as if occurring three times. In this way, the more important constructs can be selected earlier than the less important ones. Such a contextual adjustment can be generalized for co-occurrence of keywords. For example, if ``Hong'' is related to ``Kong'' as in the phrase ``Hong Kong'', we can set the co-occurrence adjustment of the two words to a value larger than 1. Each ``Hong'' or ``Kong'' alone will be treated as normal keywords, but if they occur together, the information content for the phrase is much higher. For words that are negatively correlated, the co-occurrence adjustment would be less than 1. The information content pi of ni is the sum of the weights of words multiplied by the context adjustment score in the unit.

(*a*) (*b*)

Figure 4. Sample multi-resolution browsing

a web environment, there exists a large number of unstructured documents. We are working on algorithms to extract the structure of a document from its content. We are also conducting experiments to measure the throughput of our system in browsing web documents when compared with traditional web browsing paradigm. We would like to obtain more user experiences in browsing web documents using our system and perhaps consider the concept of "intuition level" of each organizational unit in addition to its content level in defining the transmission order.

We are also integrating the concept of information content with search engines. The notion of relative information content is a good indicator for ranking documents identified by a search engine. A search engine could thus, define a more semantics-oriented ranking on the identified documents. We are working on a mechanism which allows a user to feedback relevance based on the notion of information content.

References

[1] M. Ackerman, D. Billsus, S. Gaffney, S. Hettich, G. Khoo, D. Kim, R. Klefstad, C. Lowe, A. Ludeman, J. Muramatsu, K. Omori, M. Pazzani, D. Semler, B. Starr, and P. Yap. Learning Probabilistic User Profiles. *AI Magazine*, 18(2):47–56, 1997.

[2] R. Alonso and H. Korth. Database System Issues in Nomadic Computing. In *Proceedings of the ACM SIGMOD Conference*, pages 388–392, May 1993.

[3] R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell. WebWatcher: A Learning Apprentice for the World Wide Web, 1995.

[4] M. Balabanovič. An Adaptive Web Page Recommendation Service. Technical Report SIDL-WP-1996-0041, Stanford University, 1996.

[5] N. Belkin and W. Croft. Information Filtering and Information Retrieval: Two Sides of the Same Coin? *Communications of the ACM*, 35(12):39–38, Dec. 1992.

[6] T. Bray, J. Paoli, and C. Sperberg-McQueen. EXtensible Markup Language (XML) 1.0. URL:http://www.w3.org/TR/REC-xml.

[7] B.Y.L. Chan, A. Si, and H.V. Leong. Cache Management for Mobile Databases: Design and Evaluation. In *Proceedings of the 14th International Conference on Data Engineering*, pages 54–63, Feb. 1998.

[8] T. Imielinski and B. Badrinath. Mobile Wireless Computing: Challenges in Data Management. *Communications of the ACM*, 37(10):18–28, Oct. 1994.

[9] H.V. Leong, D. McLeod, A. Si, and S.M.T. Yau. Multiresolution Transmission and Browsing in Mobile Web. In *Proceedings of the CIKM'98 Workshop on Web Information and Data Management*, pages 13–16, 1998.

[10] H.V. Leong and A. Si. Database Caching over the Air-Storage. *The Computer Journal*, 40(7):401–415, 1997.

[11] M. Mauldin and J. Leavitt. Web-Agent Related Research at the CMT. In *Proceedings of ACM International Conference on Networked Information Discovery and Retrieval*, 1994.

[12] T. Parker. Mobile Wireless Internet Technology Faces Hurdles. *IEEE Computer*, pages 12–14, Mar. 1998.

[13] B. Pinkerton. Finding What People Want: Experiences with the WebCrawer. In *Proceedings of Second International WWW Conference: Mosaic and the Web*, 1994.

[14] M. Porter. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137, 1980.

[15] M. Weiser. Some Computer Science Issues in Ubiquitous Computing. *Communications of the ACM*, 36(7):74–84, July 1993.