Hierarchical Summarization for Delivering Information to Mobile Devices

Jahna Otterbacher[a,*], Dragomir Radev[b], Omer Kareem[b]

[a]*University of Cyprus, Nicosia, CYPRUS*

[b]*University of Michigan, Ann Arbor, MI 48109*

Abstract

Access to information via handheld devices supports decision making away from one's computer. However, limitations include small screens and constrained wireless bandwidth. We present a summarization method that transforms online content for delivery to small devices. Unlike previous algorithms, ours assumes nothing about document formatting, and induces a hierarchical structure based on the relative importance of sentences within the document. As compared to delivering full documents, the method reduces the bytes transferred by half. An experiment also demonstrates that when given hierarchical summaries, users are no less accurate in answering questions about the documents.

Key words: Document summarization, Handheld devices, Information delivery

1 Introduction

In recent years, the number of Internet users accessing information via handheld devices such as Personal Digital Assistants (PDAs) and Web-enabled mobile phones has been on the rise, and is projected to continue to increase. [1] [2] Indeed, the ability to access information while away from one's desktop computer is a promising addition to Web use, as many personal information needs are generated while one is

[*] Corresponding author.
Email addresses: jahna@ucy.ac.cy (Jahna Otterbacher), radev@umich.edu (Dragomir Radev), okareem@umich.edu (Omer Kareem).

[1] eTForecast, http://www.etforecasts.com/pr/pr0603.htm

[2] eMarketer http://www.epaynews.com/statistics/mcommstats.html\#44

away from the home or office (Buyukkokten et al. , 2002; Freire et al. , 2001). Recently, it has been noted that mobile computing can and should be extended to support decision making in organizations, by providing professionals with a means to access relevant information whenever and wherever needed (Yang & Wang , 2006). This might include remote access to documents on a company intranet, checking one's email while traveling, or keeping up with the latest news and financial information. All of these activities are likely to be essential to a busy professional's ability to keep up-to-date on information pertinent to decision making.

While many applications have been developed in order to support mobile Internet access, the challenges of using handheld devices in this capacity are well documented. For example, online content is typically designed to be displayed on the screen of a desktop computer, and is often confusing and difficult to navigate when viewed on a small device. While some large Web sites, such as those of the New York Times and Amazon.com, have special versions for wireless devices, this is an excessive amount of maintenance work for most Web site administrators. In addition, while improvements continue to be made in terms of the memory size and power of handheld devices, limited bandwidth still means that users face slow download speeds when trying to view Web documents on their PDAs or phones.

To address these problems, previous research has taken two main directions. The first approach involves reformatting or adapting Web pages to be more appropriate for viewing on a small screen, without altering their original content. For instance, this might be done by splitting a given page into smaller parts to be displayed one at a time (e.g. Chen & Zhang (2003); Milic-Frayling & Summerer (2002)), or by delivering only the objects of a page deemed to be most important and eliminating non-essential items such as graphics (e.g. Yin & Lee (2004)). To contrast, a second method is to actually transform the content of the documents to be more suitable for delivery to and display on a small device, as suggested by Trevor and colleagues (Trevor et al. , 2001). Following this second approach, summarization of Web pages has been introduced as a means of presenting the user with only the most important content expressed in the text on a page (e.g. Buyukkokten et al. (2001); Seki et al. (2004); Sweeney et al. (2002); Yang & Wang (2006)). Automatic summarization is the process of condensing a text, while at the same time preserving its main points (Radev et al. , 2002). This approach addresses the issue of the limited bandwidth as it reduces the amount of data that needs to be transferred to the user's mobile device. In addition, the short summaries can be viewed more easily on a small screen.

Summarization is especially appropriate for delivering Web-based content to mobile devices in the context of decision support in organizations. Systems often present too much information to users during a decision making process (Ackoff , 1967), and tasks that involve interpreting text are especially susceptible to problems of information overload (Hiltz & Turoff , 1985). To this end, text summarization can be used as a means to filter out less relevant material, presenting the user only with that deemed necessary for the task at hand. In fact, previous research suggests that the use of summaries as a proxy for full-length texts does not significantly affect one's reading comprehension (Morris et al. , 1992). For a user who needs to access online documents via a PDA or mobile phone in order to keep abreast of work-related information, summaries can be used in order to help him or her determine if a document is relevant, before spending the time and wireless bandwidth accessing the full document.

In the current paper, we present and evaluate a novel text summarization method, which is appropriate for information delivery to small, mobile devices. Our method, which will be described in detail in Section 2 , induces a hierarchical structure for an input document, by ranking its sentences with respect to salience. Only the most representative sentences are initially delivered to the user's PDA or mobile phone, so that she can decide if the document's general topic is relevant to her needs. If needed, the user may then "drill-down" to view sentences at deeper levels of the hierarchy. It should be noted that such sentences, while deemed by the summarizer as being less salient in terms of the overall topic of the document, may express finer details of the document that interest the reader, or may answer particular questions that she has. Thus, while most methods produce summaries that are either indicative (i.e. of whether or not a document is relevant to a user's information need) or informative (i.e. providing the information contained in the whole document) (Brandow et al. , 1995), hierarchical summaries can be both indicative and informative. This is because if upon seeing the initial summary the user decides the document is in fact relevant, she can then obtain more details of the document's information content at deeper levels of the hierarchy.

Other methods for summarization of Web pages for delivery to small devices have been previously investigated. For example, Seki and colleagues (Seki et al. , 2004) developed a method that classifies sentences in an input document by their functions (e.g. sentences offering background information versus those providing a key description of the document's theme), in order to rank them by importance. In contrast to our work, they focus on creating a summarizer to handle a particular genre of document - the opinion news article. In addition, they do not aim to develop hierarchical summaries but rather, their algorithm creates short, 3-sentence summaries that can be delivered to mobile phones. Another approach is "accordion summarization," introduced by Buyukkokten and colleagues (2001, 2002), in which the basic idea is to reduce the amount of data transferred to a mobile device by sending it incrementally. However, one major difference in comparison to our work is that their method relies on HTML artifacts (e.g. tags indicating paragraphs and lists) that denote the hierarchical structure of a document.

In other related work, Yang and Wang implemented a summarization method based on the fractal theory (Yang & Wang , 2006). Their approach is especially appropriate for viewing large Web-based documents, as it makes extensive use of a document's inherent structure. For instance, an HTML news document might be made up of sections, paragraphs, sentences, terms and then words. This structure is used to create a skeleton summary, to which finer details are then added. Their method has also been applied to summarizing an entire Web site, such as Yahoo! News, as they note that news sites have an extensive hierarchical structure that can be exploited by the fractal summarization technique (Yang & Wang , 2003). In contrast to the previous approaches, hierarchical summarization can be applied to any textual online document that a user wishes to access from a mobile device, as no prior assumptions are made as to the desired document's format, structure or genre.

Presently, we will introduce and evaluate our method. While Section 2 will provide details of the summarization algorithm, we will present one application of hierarchical summarization in Section 3. In particular, we have implemented the method in an email summarization and delivery system, which accesses the user's Web-based email account, summarizes each document in its inbox and delivers the summaries to the user's mobile device. In Section 4, we will present a task-based evaluation of the implemented system, which emulates the experience of a user who wishes to keep informed of current news and financial information throughout the day using her Web-enabled mobile phone. More specifically, the subjects in our study used the hierarchical summaries as well as several baseline methods, to answer questions about a set of news articles sent to the user's email account (e.g. by a news alert service). We will demonstrate that the subjects achieved better task accuracy when using the hierarchical summaries, as compared to three other summarization methods. In addition, as compared to the case where the full text of the articles are sent to the user's phone, the new method reduces the number of bytes transferred per user request by more than half. Even more promising is the finding that users took no longer to complete the tasks and were just as accurate as they were in the case where they were sent the full text articles. In other words, our results concur with those of previous researchers in that automatically-produced summaries can provide just enough information to users, so that they can make decisions as effectively as if they had read the full text of the document (Morris et al. , 1992).

2 Hierarchical Summarization

In this section, we describe our hierarchical summarization algorithm. Figure 1 illustrates the architecture of our summarization method, which is a three-step process involving document preprocessing, the scoring of sentences with respect to their importance to the theme of the document, and finally, the hierarchical nesting of the sentences. In building the system, we have employed the MEAD summarization environment (Radev et al. , 2004). The input to the summarizer is a Web-based document. Given the document to summarize, the textual content is extracted from it, and is segmented into sentences. As described in (Radev et al. , 2004), this is done via the use of regular expressions. Next, the sentences are assigned salience

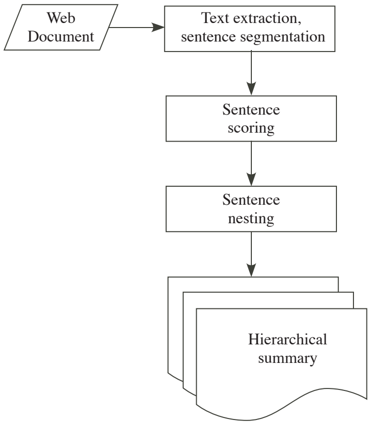


Fig. 1. Summarization architecture.

scores and are nested hierarchically based on these scores. These two steps are described in detail below. It should be noted here that our summarizer treats only the textual content of an input document; the treatment of graphics, sound files and other data found on a Web page is beyond the scope of our current work.

2.1 Sentence scoring

In the second step of the process, a salience score is computed for each sentence in the input document. Here, we perform the scoring algorithm used in the centroid-based summarization method, which has been shown to produce generic (i.e. not query-sensitive) summaries of a similar quality to those created manually by humans (Radev et al. , 2004a). In this procedure, the final score for a given sentence S_i, is the sum of three feature-based scores of S_i: the centroid value, positional value, and its overlap with the first sentence in the document. The final scores of the sentences are used to place them within the hierarchical summary, as will be described in 2.2.

The centroid value of S_i, C_i, represents the importance of the sentence, in terms of how well it represents the central topic of the input document to be summarized. In particular, C_i quantifies the extent to which the sentence contains words that are statistically important to the document. While the centroid value has often been used in creating summaries from a set of topically related documents (i.e. in producing multi-document summaries, as detailed in (Radev et al. , 2004a)), the centroid method is also appropriate for creating summaries of single documents. In this case, an input document is viewed as a set of sentences on possibly different topics.

To obtain the centroid value for each sentence in the document, the document centroid is first created. This is done by representing the document as a weighted vector of the words in it. Each word (or element) in the vector has a weight corresponding to its TF*IDF value, where TF is the term frequency of the word in the input document, and IDF (inverse document frequency) measures the proportion of all documents in a large collection of documents that contain the word (Salton & Buckley , 1988). In other words, content words that are important in describing the document's topic will have high TF*IDF values as compared to more common words that do not convey as much topical information about the document. As detailed in (Radev et al. , 2004a), the centroid for a document to be summarized consists of all words in the document that have a TF*IDF score above a pre-defined threshold, thus eliminating function words that do not contain topical information (e.g. articles, prepositions, etc.) The weight, or centroid score, of each word in the document, $C_{w,i}$, is its respective TF*IDF score. In scoring a given sentence S_i, its centroid value is the sum of the centroid values, $C_{w,i}$ of all of the words it contains:

$$C_i = \sum_w C_{w,i}$$

The next sentence characteristic used in scoring, the positional value of S_i, P_i, is based on its position in the input document to be summarized. More weight is given to sentences that appear earlier in the document than those that appear later. It should be noted here that this implementation of the position score is a design choice based on the fact that we are currently summarizing news articles, which are typically written using the "inverse pyramid" structure, in which more important information is given to the reader first, followed later by finer-grained details of a story (Mitchell & West , 1996). In cases where this characteristic is not likely to hold true, the position parameter could be modified or eliminated from the sentence scoring mechanism. For example, in the case of summarization of personal email, where we might expect the first one or two sentences of the text to express a salutation rather than the main point of the communication, we should tune the sentence position score to take this into account.

In the current implementation, the first sentence in the document is assigned the score, C_{max}, which corresponds to the centroid value of the highest-ranking sentence. The positional value of the sentences are computed as follows, where there are n sentences in the document:

$$P_i = \frac{(n-i+1)}{n} * \max_i(C_i)$$

Next, the first-sentence overlap value, F_i, quantifies the similarity of S_i to the first sentence in the document. The first sentence in a text is likely to convey information about its main theme or topic (i.e. a "topic sentence"). S_i and the first sentence of the document are represented as k-dimensional vectors of words, where k is the number of words in the document's centroid (i.e. k is the number of content words in the document with a TF*IDF score over the predetermined threshold). The value at each position i in the sentence vector is the number of occurrences of the given word, w_i, in the sentence. The overlap value for S_i is then the inner product of the two sentence vectors:

$$F_i = \vec{S_1}\vec{S_i}$$

Thus we are comparing the overlap of the content words in the two sentences. Again, like the positional value, the first-sentence overlap value is a quantity that can be tailored depending on the genre of the text to be summarized. For example, in the case of summarizing personal email, the subject line might be substituted in place of the first sentence, in order to identify sentences that are likely to contain information important to the central topic of the email.

Finally, the salience score of S_i is the sum of the three feature scores:

$$SCORE(S_i) = C_i + P_i + F_i$$

2.2 Hierarchical nesting

In the third stage of processing, the salience scores of the sentences are used to induce a hierarchical structure. In addition, as will be illustrated, the positioning of the sentences in the source document is respected in order to promote cohesion in the resulting summary. A tree is constructed from all of the sentences such that its root is the sentence with the highest salience and, given any sentence node with salience l at depth d, all sentences above that depth have a salience greater than l, while the scores of the rest of the sentences are below l. As will be shown, lower ranking sentences are simply hidden at each level, so that the order of presentation of the sentences in the source document is preserved.

As explained in Section 1, one key feature of hierarchical summaries is that they can be both indicative and informative with respect to the summarized document and the user's information need. Presently, we have implemented the sentence nesting algorithm such that at the first level, the top four sentences are shown to the user. In other words, the user first sees the four sentences deemed as being the most representative of the document's overall topic. This can help the user in deciding if the topic is interesting to her. If the topic is indeed relevant to the user's need, the lower-ranking sentences can then be viewed by going deeper into the hierarchy (i.e. by "expanding" the summary). Specifically, in the current implementation, at each of the lower levels, the next three best sentences are shown. It should be noted that the number of sentences to be shown to the user at each level can be varied according to the screen size of the mobile device to be used and/or with respect to how conservative a user wishes to be in terms of using available bandwidth to transfer text to her phone or PDA.

Figure 2 illustrates the hierarchical summary that is produced for a document containing 20 sentences, whose salience scores and relative rankings are shown in Table 1. In addition, the full text of the article summarized can be found in Figure 3. As can be seen in Figure 2, the four most salient sentences are numbers 1, 2, 3 and

10, which are shown at the initial level of the hierarchy. Sentences 4-9, and 11-20, have been hidden from the user. To create the second level of the hierarchy, the algorithm finds the next three highest-ranking sentences between sentences 4-9 (4, 7, and 8) and between sentences 11-20 (11, 13, and 14). Using this procedure, deeper levels are added to the hierarchy until all sentences have been included.

Table 1
Sentence salience scores used by the nesting mechanism.

Sentence	Score	Rank	Sentence	Score	Rank
1	1.97	1	11	0.95	9
2	1.42	2	12	0.80	12
3	1.34	3	13	1.17	6
4	0.95	10	14	0.97	8
5	0.73	16	15	0.95	11
6	0.78	13	16	0.77	15
7	0.99	7	17	0.65	18
8	1.19	5	18	0.71	17
9	0.50	19	19	0.36	20
10	1.32	4	20	0.77	14

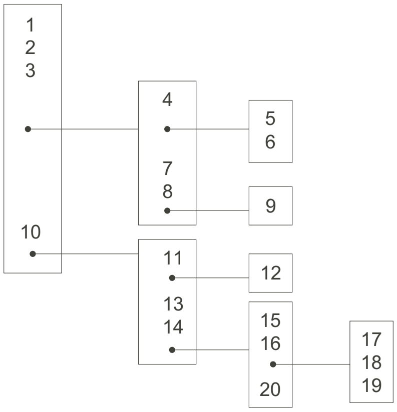
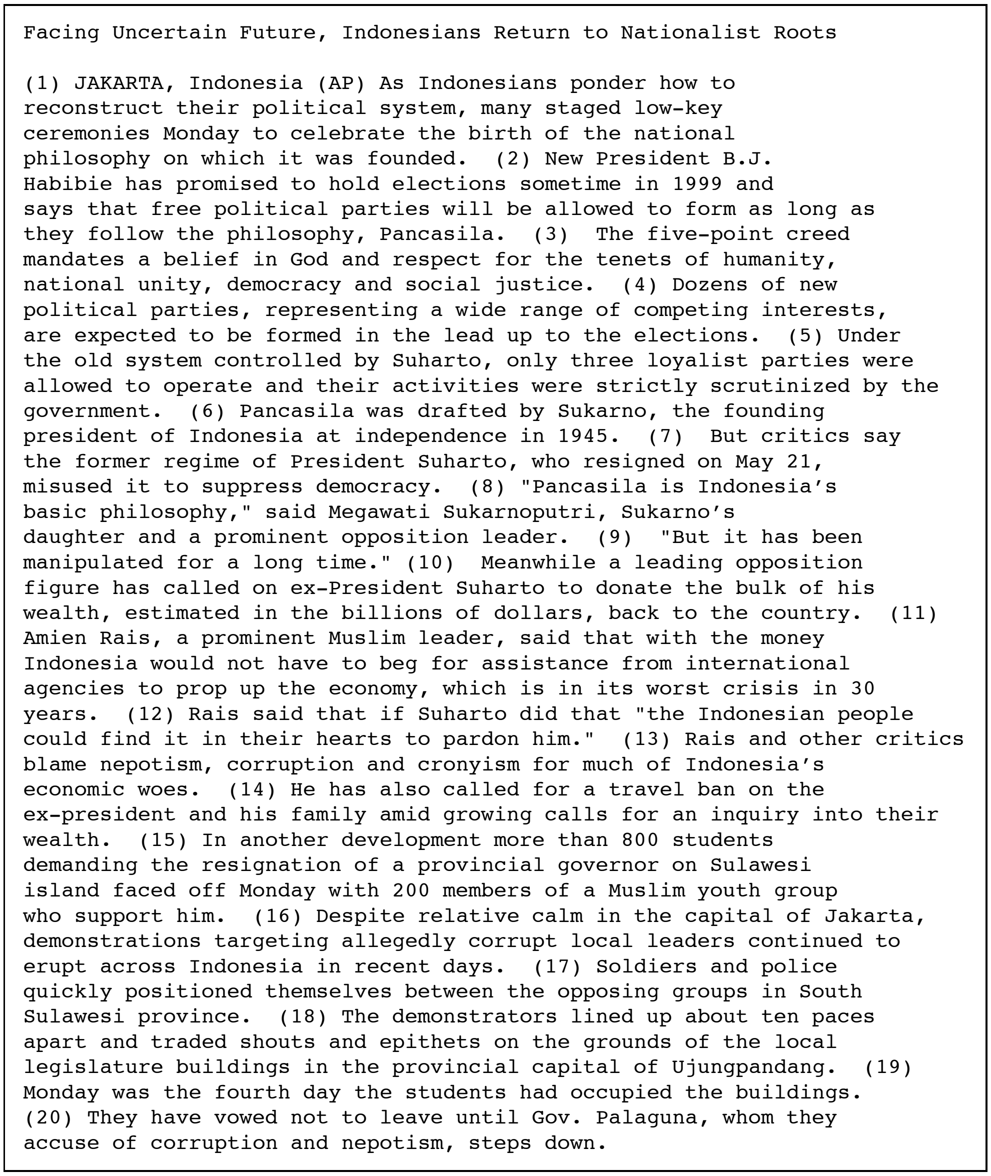


Fig. 2. Hierarchical sentence nesting.

The idea behind hierarchical summarization is that the user is first shown the most representative sentences of a document, in order to get the gist of it. If she finds the initial summary interesting or relevant to her information need, she may drill down the finer details of the document by expanding the summary. In other words, the initial 4-sentence summary can be viewed as an indicative summary, which can help a user in deciding if the given article is interesting enough to invest more of her resources in order to view more of the document. Thus, the motive is to save time, bandwidth and screen space by delivering and displaying the most salient information first, while at the same time giving the user the opportunity to view more of the document if desired.



```
Facing Uncertain Future, Indonesians Return to Nationalist Roots

(1) JAKARTA, Indonesia (AP) As Indonesians ponder how to
reconstruct their political system, many staged low-key
ceremonies Monday to celebrate the birth of the national
philosophy on which it was founded.  (2) New President B.J.
Habibie has promised to hold elections sometime in 1999 and
says that free political parties will be allowed to form as long as
they follow the philosophy, Pancasila.  (3)  The five-point creed
mandates a belief in God and respect for the tenets of humanity,
national unity, democracy and social justice.  (4) Dozens of new
political parties, representing a wide range of competing interests,
are expected to be formed in the lead up to the elections.  (5) Under
the old system controlled by Suharto, only three loyalist parties were
allowed to operate and their activities were strictly scrutinized by the
government.  (6) Pancasila was drafted by Sukarno, the founding
president of Indonesia at independence in 1945.  (7)  But critics say
the former regime of President Suharto, who resigned on May 21,
misused it to suppress democracy.  (8) "Pancasila is Indonesia's
basic philosophy," said Megawati Sukarnoputri, Sukarno's
daughter and a prominent opposition leader.  (9)  "But it has been
manipulated for a long time." (10)  Meanwhile a leading opposition
figure has called on ex-President Suharto to donate the bulk of his
wealth, estimated in the billions of dollars, back to the country.  (11)
Amien Rais, a prominent Muslim leader, said that with the money
Indonesia would not have to beg for assistance from international
agencies to prop up the economy, which is in its worst crisis in 30
years.  (12) Rais said that if Suharto did that "the Indonesian people
could find it in their hearts to pardon him."  (13) Rais and other critics
blame nepotism, corruption and cronyism for much of Indonesia's
economic woes.  (14) He has also called for a travel ban on the
ex-president and his family amid growing calls for an inquiry into their
wealth.  (15) In another development more than 800 students
demanding the resignation of a provincial governor on Sulawesi
island faced off Monday with 200 members of a Muslim youth group
who support him.  (16) Despite relative calm in the capital of Jakarta,
demonstrations targeting allegedly corrupt local leaders continued to
erupt across Indonesia in recent days.  (17) Soldiers and police
quickly positioned themselves between the opposing groups in South
Sulawesi province.  (18) The demonstrators lined up about ten paces
apart and traded shouts and epithets on the grounds of the local
legislature buildings in the provincial capital of Ujungpandang.  (19)
Monday was the fourth day the students had occupied the buildings.
(20) They have vowed not to leave until Gov. Palaguna, whom they
accuse of corruption and nepotism, steps down.
```

Fig. 3. Text of article summarized in the hierarchical sentence nesting example.

3 Application: A News Delivery System

In order to evaluate the effectiveness of our summarization method, as well as to demonstrate its potential in terms of supporting decision making when away from one's desk, we have implemented a news summarization and delivery system that works with the user's Web-based email account. The current system was implemented using the DeckIt Wireless Application Protocol (WAP) mobile phone emulator[3], as well as a Google Gmail account. As previously mentioned, we address the scenario in which a user wants to keep current on news and financial information throughout the day, by accessing Web documents on her mobile phone. We

[3] http://www.wats.ca/tools/deckit-1.2.4.exe

assume the the user subscribes to a service in which news and financial articles are sent to her Gmail account, and she checks the account periodically while away from her desk.

Figure 4 gives an overview of the implemented system. Periodically, the system's mail daemon connects to the Gmail account's inbox and caches new email into the system's disk. The detection of new mail thus starts the summarization process previously described in Figure 1. Once the summaries of the documents are ready, they are sent to the mobile phone emulator, over the WAP protocol.

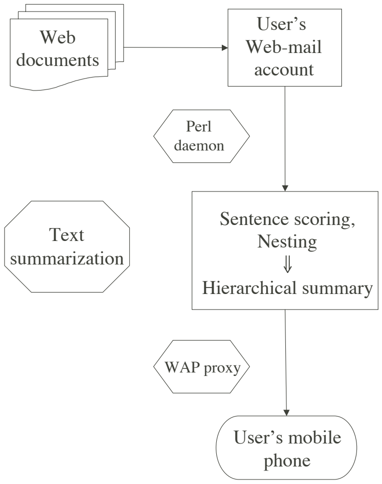


Fig. 4. System architecture: Summarization and delivery to WAP-enabled device.

Figures 5 and 6 illustrate how the user interacts with the system. As shown in Figure 5, what the user views first on the phone's screen is the list of documents that are in the email account's inbox. In particular, the documents are listed by number (i.e. in the order in which they were received) with their titles. In this case, the user wishes to view news article 6, which is the same article used in the examples in Section 2.2. In the middle panel of Figure 6, the user has opened up the first level of the summary for article 6, in which sentences 1, 2, 3 and 10 are displayed. Finally, the right-most panel shows what the user sees when sentences have been hidden in the hierarchical summaries. In this case, sentences 4-9 have been hidden, but can be expanded by clicking on the corresponding highlighted area. This action would take the user to the second level or depth of the hierarchy for article 6, in which sentences 4, 7 and 8 would be delivered to her phone.

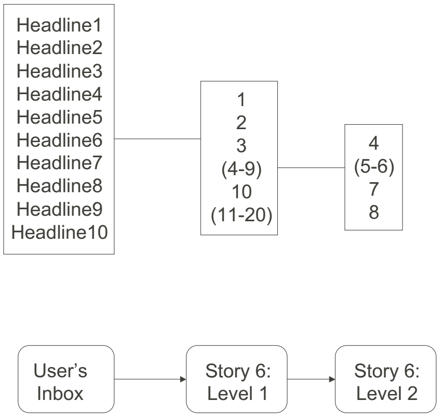
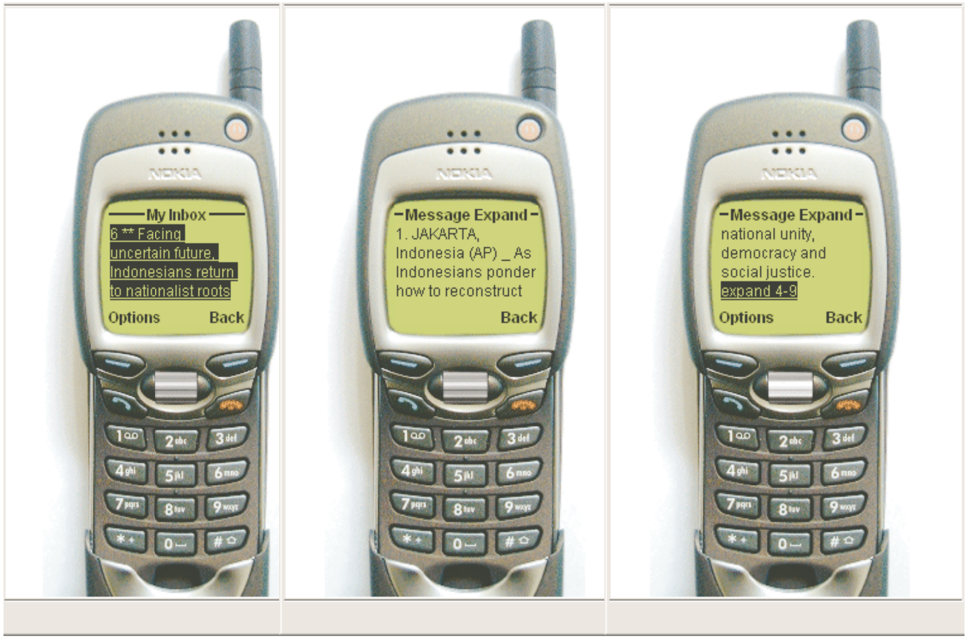


Fig. 5. Hierarchical structure of summary shown in Figure 6.



6 ** Facing uncertain future, Indonesians return to nationalist roots

1. JAKARTA, Indonesia (AP) _ As Indonesians ponder how to reconstruct

national unity, democracy and social justice.
expand 4-9

Fig. 6. System interface.

4 Experiment

We conducted a task-based user experiment in order to evaluate the effectiveness of hierarchical summarization for delivering information to a user's small, wireless device. In particular, we wanted to evaluate our method, as implemented in the Web-mail summarization and delivery system described in Section 3, as to how well it facilitates the finding of desired information, when using one's mobile phone rather than desktop computer. Therefore, we created an experiment in which subjects were asked to complete several tasks that involved both finding and comprehending textual documents, with the use of automatically-produced summaries sent to a mobile phone. In other words, we have used the "simulated information needs" approach to evaluating our summarization method (Borland & Ingwersen , 1997), in which realistic information seeking tasks are given to subjects in a controlled environment.

As will be detailed in 4.1, for our experiment, we designed a task similar to that of Morris and colleagues (1992), which is designed to measure a user's reading comprehension of a set of articles that have been summarized, and in which users are asked to find the answers to key factual questions surrounding the articles. In addition to Morris and colleagues, others have also evaluated summary quality and usefulness based on how well they facilitate users in finding facts (e.g. (White et al. , 2001)) and answers to questions (e.g. (Seki et al. , 2004)) from Web-based documents. This framework allowed us to assess the effectiveness of our hierarchical summarization method in a realistic setting.

4.1 Tasks

We created five unique sets of news documents, consisting of 10 Associated Press (APW) articles each. The 10 articles in a given set were published on the same day, and were emailed to a Gmail account that was established for our experiments. Each set of articles collected represents a snapshot of current events at a particular point in time, and included both world news stories as well as financial updates. In the experiments, subjects used our system to find answers to a set of questions about the set of news stories they were given. In particular, we created a task for each of the five document sets, with each task involving 10 questions (one question about each article).

We used multiple choice questions in which there were five possible answers, but only one correct answer, for each question. An example of a document set (i.e. the headlines of the 10 documents mailed to the user in one task) is shown in Figure 7. The questions comprising each task concerned the central facts about the stories or events described in the respective articles. For instance, the questions used in the task for the document set shown in Figure 7 are given in Figure 8. In addition, the possible answers are shown for the first four questions. For all questions in all of the five tasks, the answers were reported in their respective articles, although they could not be answered by reading the title of the article alone. Therefore, it was not necessary for subjects to use previous world knowledge or reasoning in order to answer the questions.

```
1.  Vietnamese journalist awarded for devotion to free press
2.  India's government faces budgetary woes
3.  Malaysia Finance minister: Bad stats won't change growth forecast
4.  TV, telephone, computer developments in Asia discussed
5.  Australia clinches first wheat sale to Egypt mill project
6.  Papua New Guinean leader to be first to greet Habibie
7.  Protege of scandal-plagued president heads for runoff with rival
8.  BC Britain Opening Gold
9.  Dollar rises, stocks plunge in Tokyo trading prices
10. India denies it has plans for another nuclear test With Pakistan-India
```

Fig. 7. Example document set.

```
1. Who is Malaysian prime minister Mahathir's closet economic adviser?
a.  Daim Zanuddin
b.  Jalam Semarak
c.  Anwar Ibrahim
d.  Tuanku Syed Sirajuddin
e.  Pak Lah

2. How much wheat does Egypt import?
a.  450 tons
b.  4,000 tons
c.  600,000 tons
d.  6,500,000 tons
e.  120,000,000 tons

3. What is Kwinana?
a.  A town in Gambia
b.  The capital of the Comoros
c.  A township in South Africa
d.  A port in Australia
e.  A large city in Sierra Leone

4. Which stock index was down by 2.24 percent?
a.  The NASDAQ
b.  The FTSE
c.  The Nikkei
d.  The TOPIX
e.  The Tokyo Stock Price Index

5. What was Doan Viet Hoat's occupation in 1976?
6. Who is Yashwant Sinha?
7. Where is Port Moresby?
8. Where is Irian Jaya?
9. What percentage of the vote in the Columbian presidential elections
   did the Liberal party win?
10. What is the name of a Japanese Car Producers' Organization?
```

Fig. 8. Example set of 10 factual questions.

4.2 Treatments

In completing a given task, a subject was assigned to one of six treatments (or system settings). In addition to the hierarchical summarization setting, we included settings at the two extremes: the full text setting, in which the unaltered, original articles were sent to the user, as well as the setting in which nothing was given to the subjects other than the task questions. This control setting accounts for the possibility that the questions themselves may provide some information about the news stories to the subjects, or that the subjects may know some of the facts sur-

Table 2
The six experimental treatments.

Treatment	Description
Full Text	Full, unaltered text of each news article in the inbox
Hierarchical Summary	Nested summary showing the top 4 sentences, followed by the next 3 ranking sentences, for each article
Top 20% Summary	Displays the top 20% of sentences for each article, in the order in which they appear in the source document
Lead-based Summary	Shows only the first 4 sentences of each article
Random Summary	20% of the sentences in each article are chosen at random for inclusion in the summary
No Summary	No news articles or summaries given

rounding the news stories. Finally, we also administered three other summarization methods, which are commonly used as baselines in evaluating text summarization systems (e.g. in the Document Understanding conference (Over & Yen , 2003)) - a top 20% summary (the highest-ranking sentences by salience scores, as described in Section 2), a lead-based summary as well as a randomly-produced summary. All six treatments used in the experiment are described in Table 2.

Note that the system is used in the first five settings, with nothing being presented to the user in the control ("no summary") setting. In addition, in the five treatments in which the system is used, the user sees the same initial display on the mobile phone, regardless of the summarization setting. In other words, the user first sees the email inbox, which shows a list of the 10 headlines of the articles in it, as depicted in the left-most panel of Figure 6. Therefore, Table 2 describes what the subject sees after selecting one of the news story headlines from the inbox.

4.3 Experimental design

A total of 39 subjects was used in the experiment. They were recruited through an email sent to students studying information and computer sciences at the University of Michigan. All subjects self-reported as native or near-native English speakers who were experienced Web users. Finally, they were paid for their participation in the study.

Although they were encouraged to complete all five of the tasks, the subjects were not required to do so (due to university research policies). Therefore, the researchers made sure that each of the five tasks and six treatment settings were assigned approximately equally often in the experiments. A balanced, incomplete block design was used, and the counts of each of the 30 possible document set-treatment pairings are shown in Table 3. In addition, the treatment and document set orderings were varied in order to prevent learning effects. Here, one should note that while the varying of the document set and treatment pairings helps in reducing potential topic effects, user effects may be significant in our study, as is often the case in information retrieval user studies. For example, even when experimental subjects have a similar level of search experience, differences in innate abilities (e.g. in reading comprehension or in overall cognitive processing) may lead to differences in performance (Turpin & Scholer , 2006).

Table 3
Counts of the document set and treatment pairings.

Docset	T1	T2	T3	T4	T5	T6	Total
1	5	5	3	5	6	6	30
2	5	4	7	5	7	3	31
3	5	2	11	4	5	5	32
4	5	5	1	2	4	8	25
5	2	7	4	8	2	3	26
Total	22	23	26	24	24	25	144

Before the experiments, the subjects were not given any information about the system that they would be using. They were informed that they would be participating in an information retrieval study and that its purpose was to examine how people search for information using a Web-enabled mobile phone. Finally, the subjects were told to answer the questions in each task as accurately as possible and were given unlimited time to complete the tasks.

5 Variables Studied and Research Questions

In comparing the users' performance on the information-seeking and comprehension tasks across the six experimental treatments, we examined the time taken to complete a task (recorded in minutes and seconds), as well as task accuracy (i.e. proportion of questions correctly answered, in which each question is either correct or incorrect). These are commonly used measures in extrinsic, or task-based evaluations of text summarizers (Hand , 1997). In addition, they have also been used to quantify the extent to which subjects achieve satisfactory reading comprehension through the use of summaries (Morris et al. , 1992). Furthermore, we obtained the number of requests made by the user, which also corresponds to the number of mouse clicks (or hits) in this case, as well as the total number of bytes transferred while completing a task. This information was obtained from the log file of each user's session (completing one task using one system setting). We then computed for each session, the number of bytes transferred per user request, in order to compare the efficiency of each of the methods tested. This measure gives us an idea of how much data has to be transferred to and displayed on the user's wireless device each time he or she interacts with the system, in order to complete the search task at hand.

The means of the three response variables across the six settings are shown in Ta-

Table 4
Mean time to task completion, task accuracy and bytes transferred per click under each setting.

Setting	Time (min.)	Task accuracy	Bytes per click
Full text	19.5	0.94	2674.5
Hierarchical summary	17.5	0.83	1206.0
Top 20% summary	16.1	0.63	1175.4
Lead-based summary	15.2	0.68	1295.4
Random summary	12.7	0.59	1208.2
No summary	2.9	0.32	0

ble 4. The subjects were most accurate on the information-finding tasks when using the setting in which they were shown the full text of the news articles, with an average task accuracy of 0.94. They also took more time to complete the tasks (an average of 19.5 minutes) than they did when using the article summaries. This finding is not very surprising, as in the full text case, the answers to the questions will always be available to the user, such that it is simply a matter of taking the time to find them. However, as will be shown in Section 6, the differences in time and accuracy are not significant between the full text setting and that in which subjects used hierarchical summaries to complete the tasks.

Another expected finding is that all of the summarization methods reduce the data transferred per user request, by more than half as compared to the full text setting. This is intuitive since, when using the summarization techniques, the goal is to prioritize information by ranking the sentences according to salience and to display the sentences incrementally in rank order. Finally, we can see that in the "no summary" treatment, where subjects answered questions about a document set without access to the documents or their summaries, the accuracy is very low (an average of 0.32). Therefore, there is no evidence that the questions themselves contain too much information about the news stories and we are not concerned about the task being trivial.

In the next section, we will concentrate on answering three research questions using the data from the user study:

1. **Are there significant differences between the five treatments (systems) when the effects of task difficulty are controlled?**

The means of the three response variables, time to task completion, task accuracy and bytes transferred per hit, which are shown in Table 4, appear to differ between the five systems. In addition, in assigning subjects to a given task (i.e. set of documents and questions to answer) and setting, we tried to ensure an approximately even distribution of task-setting pairing. Nonetheless, we want to investigate the effect of the five system treatments on the three response variables when the possible effects of the task are controlled. For example, it may be the case that some tasks were more difficult than others. Likewise, it could be possible that certain task and system combinations resulted in longer task completion times or lower rates of accuracy.

2. **Are there any significant differences in task performance and efficiency between the hierarchical summarization setting and the full text setting?**

If we establish, in investigating our first research question, that there is a significant system effect on the three response variables, then we should make pairwise comparisons between the five systems. In particular, as shown in Table 4, the highest mean task accuracy (0.94) occurs when subjects use the full text of the news documents to complete the tasks. To contrast, the accuracy when using the hierarchical text summaries of the news articles is slightly less, at 0.83. After that, we see a drop off, as the system with the next best accuracy, the lead-based summary setting, has a mean accuracy of only 0.68. Therefore, we will compare the hierarchical summary case versus the full text setting in order to see if the differences between them are statistically significant.

3. **Are there significant differences between the hierarchical summarization setting and the other three summarization methods?**

Finally, we wish to compare the three response variables between the hierarchical summarization setting and the three other summarization methods. As mentioned previously, these three methods are commonly viewed as baseline systems. As can be seen in Table 4, all four of the summarization methods reduce the number of bytes transferred per hit, as compared to the full text case. Therefore, we want to investigate whether the new, hierarchical summarization method offers any significant advantages over the baseline methods in terms of the users' performance on the tasks.

6 Analysis

Below, we analyze the data collected from our user study in order to address the three research questions put forward in the previous section.

6.1 Setting effect when task is controlled

In order to examine if there is an overall setting (or system) effect, when controlling for the task administered to the subjects, we conducted an analysis of variance (ANOVA) for each of the three response variables, which were all approximately normally distributed. First, we removed the cases where subjects were given only the tasks with no source articles or summaries (a control setting), in order to consider only the differences between the five systems (where either the full text doc-

Table 5
P-values for predictors Setting, Task and their interaction for ANOVAs on each response variable.

Response variable	Setting	Task	Setting*Task
Time	0.0033	0.0944	0.3714
Accuracy	0.0000	0.0549	0.0756
Bytes per click	0.0000	0.0023	0.4222

Table 6
P-values for the differences in the response variables between the full text and hierarchical summarization settings.

	Difference	P-value
Time (min.)	2.0	1.000
Accuracy	0.11	0.645
Bytes per hit	1468.5	0.000

uments were given to the user or one of the four types of summaries). In each ANOVA, the predictors were the setting used, the task/document set used, and the interaction between the given setting and task. For the ANOVAs on each of the three response variables, Table 5 shows the p-values of the three predictor variables.

As can be seen in the table, the setting effect is highly significant in all three ANOVAs, even when controlling for the effect of task. In fact, at the 5% significance level, the effect of the task assigned was significant in only one case, when the number of bytes per click is the response variable. Likewise, at this level, the interaction effect between the task and the system assigned is not significant for any of the response variables. (However, at the 10% level, the interaction is significant in the ANOVA of the response variable “accuracy.”)

Therefore, we can conclude that there are significant differences between the five system settings in terms of the average time to complete the information-seeking task, the average task accuracy, and the number of bytes transferred per mouse click, even when we control the effects of the tasks and the interaction between the setting and task administered.

6.2 Hierarchical summarization versus the full text setting

Having established that there are significant differences between the five system settings, we can now make pairwise comparisons between them, in order to see which systems are better than others, in the context of the current task. Post-ANOVA pairwise tests can be conducted using the Bonferroni method (Neter et al. , 1990). Table 6 displays the differences in the average task completion times, task accuracy and bytes transferred per hit, between the full text setting and that in which users were shown hierarchical summaries of the documents in the email inbox. In addition, the corresponding Bonferroni-corrected p-values are shown.

Table 7
Significant differences and their p-values in response variables between the hierarchical and baseline summarization settings.

Comparison	Response variable	Difference	P-value
Hierarchical vs. Top 20%	Accuracy	0.20	0.0010
Hierarchical vs. Lead-based	Accuracy	0.15	0.0660
Hierarchical vs. Random	Time	-4.8	0.0300
	Accuracy	0.24	0.0000

We can see that the differences between the two systems with respect to the average time taken to complete the task and the task accuracy are not statistically significant, having large p-values of 1 and 0.6, respectively. To contrast, the difference in the mean number of bytes transferred is highly significant, with a p-value of 0. The interpretation of these findings is that there is no evidence of significant performance differences on the task of finding answers to questions about a set of news stories between the two settings. However, the use of hierarchical summarization in delivering newsworthy information to a user's mobile phone reduces the number of bytes transferred to the wireless device each time the user interacts with the system.

6.3 Hierarchical summarization versus the baseline methods

The post-ANOVA pairwise comparison tests were also used to examine the differences in the response variables between the hierarchical summarization setting and each of the three baseline summarization methods. The statistically significant differences are shown in Table 7 along with their corresponding p-values. It should be noted that the difference in accuracy between the hierarchical and the lead-based summarization methods is not significant at the 5% level, but only at the more lenient significance level of 10%.

As can be seen, users achieved better task accuracy when using the hierarchical summaries, as compared to the other three summarization methods. On average, the users took 4.8 minutes less to complete the tasks when using the random summaries as compared to the hierarchical summaries. However, the low accuracy achieved using randomly-created summaries (average accuracy of 0.59 as compared to 0.83 in the hierarchical summary setting) confirms one's intuition that the randomly generated summaries are of a relatively poor quality. Therefore, we suspect that the shorter task completion times might reflect users "giving up" on a search task if they are unable to find the answers to questions after exerting significant effort. In conclusion, while the hierarchical summarization method does not offer an advantage over the baselines in terms of the time taken to complete the information-seeking tasks, the users achieved significantly better task accuracy using the new method.

7 Discussion

As previously discussed in the introduction, text summarization, as a means to format Web documents for delivery to a small, handheld device, is not a new proposal. In particular, the methods put forward by both Buyukkoten and colleagues (2001, 2002) and Yang and Wang (2006) are similar in spirit to ours, in that the idea is to reduce the amount of text that is delivered to and viewed by the user at each point in time. While both of these previous approaches use the structure inherent in Web documents, either by using the information conveyed by HTML tags (Buyukkokten et al. , 2001, 2002), or the inherent structure of large documents or Web sites (Yang & Wang , 2006), we have put forward a method that assumes nothing about the format of the document to be summarized. This allows us to treat a range of documents that a user might want to view (e.g. news articles, an email message, a report posted by a colleague).

As noted by Yang and Wang, it has been suggested that a "tree view" (Buyukkokten et al. , 2001) or a "hierarchical view" (Mani , 2001) is desirable for displaying text summaries on handheld devices. However, summarization methods based on the sentence extraction method treat a document as being a sequence of sentences, with a flat structure. Our current work illustrates one approach to this problem. While true that we have induced a hierarchical structure on an input text document, rather than using the existing information available from HTML or other artifacts, the results of our extrinsic evaluation indicate that summaries are useful for those who access them on a handheld device. Also, as previously mentioned, this approach allows us to summarize texts in any format.

One goal of the current work was to show that hierarchical summaries not only effectively transform textual data for delivery to a wireless device, but also that they can facilitate decision making by Web users when away from a desktop computer. To this end, we presented a task-based user experiment, which illustrated how the summaries help users find relevant information in textual documents. In contrast to the previous research mentioned above, Yang and Wang (2006) conducted smaller user evaluations of their fractal summarization method. However, their study's focus was different than ours, in that subjects were asked to subjectively quantify the quality of the automatically-produced summaries, rather than to perform information-seeking and comprehension tasks using the summaries. Buyukkokten and colleagues conducted a small user study in which subjects performed tasks on a PDA emulator, but focused more on Web-browsing rather than on decision making.

While not specifically considering the problem of delivering documents to handheld devices, other research has also demonstrated the value of summarization as a tool for deciding which information is important in a text or set of texts. In the work of Morris and colleagues (1992), subjects completed multiple choice questions, taken from a GMAT exam, that gauged subjects' level of comprehension of the corresponding texts. They found no significant difference in reading comprehension (task accuracy) between the case in which subjects were given the full texts, versus that in which they were given summaries (human-constructed summaries, or 20% to 30% automatically-produced extractive summaries). To contrast, McKeown and colleagues (2005) investigated the use of automatically-produced summaries of news articles in helping subjects create a report or overview of the news. Their results showed that using the system-produced summaries resulted in better reports than did the use of the original documents only, or the lead-based summaries. In sum, the results of our experiment concur with those of others that have found that text summarization is an effective tool. In particular, summarization can reduce the amount of text users must read, while at the same time not hindering their comprehension of the key ideas expressed in the text.

8 Conclusion

The use of mobile devices, such as PDAs and Web-enabled phones, to access online information has emerged as a means to complement use of the Web and Internet, by empowering users to easily access and interact with information and services instantly from anywhere they might be. Making content on the Internet available to a mobile device is a distinct and legitimate challenge, as there are many constraining factors limiting the media presented onto such small devices. As previously discussed, screen size, low bandwidth and difficult interaction can be seen as some of the main barriers to overcome. However, as much of the information demanded from the Internet is in textual form, text summarization is a means to make Web-based context more easily accessible and deliverable to the mobile user.

In the current paper, we presented a method to summarize online textual documents, by first determining which sentences are likely to be the most representative of the document, and then by nesting them in a meaningful hierarchical structure. Given that the method summarizes plain text, and does not make any assumptions about an incoming document's format, it can be implemented and used in a number of ways. Presently, we showed how hierarchical summarization can be paired with a user's Web-based email account, in order to deliver short summaries to a mobile device. While the current application featured the summarization of news articles sent to a user's account, emulating the experience of a user who wants to keep up on current events while away from the home or office, the method is certainly appropriate for a wide range of applications in which a user needs remote access to information expressed in Web-based documents. For example, access to financial documents for mobile devices can support decision making in organizations competing in a fast-paced economy (Yang & Wang , 2006). Similarly, in the domain of healthcare, doctors working in the field can use mobile devices to access documents in order to answer urgent questions that come up on the job (Mendelsohn , 2001). Regardless of the domain of application, what is common is the need of mobile users to have fast and convenient access to Web-based information that can support their work while away from a desktop computer.

In conclusion, the results of the task-based user experiment presented suggests that hierarchical text summarization is an effective means to support users in accessing and comprehending online textual documents, which in turn supports decision making. In particular, it saves users effort by providing them with a means to decide if a given textual source is likely to contain the information they need, before spending the time downloading the full document to the remote device. Therefore, the limiting factors associated with using handheld devices are significantly reduced.

Acknowledgments

This work was partially supported by the U.S. National Science Foundation under the following grant: 0329043 “Probabilistic and Link-based Methods for Exploiting Very Large Textual Repositories” administered through the IDM program. All opinions, findings, conclusions, and recommendations in this paper are made by the authors and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank the members of the CLAIR research group at the University of Michigan, as well as the two anonymous reviewers for their feedback and comments on this work.

References

Ackoff, R. L. (1967.) Management Misinformation Systems. Management Science, 14(4), (pp. 147-156).

Borlund, P. & Ingwersen, P. (1997.) The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. Journal of Documentation, 53(3), (pp. 225-250).

Brandow, R., Mitze, K. & Rau, L. (1995). Automatic Condensation of Electronic Publications by Sentence Selection. Information Processing and Management, 31(5), (pp. 675-685).

Buyukkokten, O., Garcia-Molina, H. & Paepcke, A. (2001.) Accordion Summarization for End-game Browsing on PDAs and Cellular Phones. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2001), Seattle, Washington, (pp. 213-220).

Buyukkokten, O., Garcia-Molina, H. & Paepcke, A. (2001.) Seeking the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices. In Proceedings of the 10th World Wide Web Conference (WWW ’01), Hong Kong.

Buyukkokten, O., Kaljuvee, O., Garcia-Molina, H., Paepcke, A. & Winograd, T.

(2002). Efficient Web Browsing on Handheld Devices Using Page and Form Summarization. Association for Computing Machinery Transactions on Information Systems (ACM TOIS), 20(1), (pp. 82-115).

Chen, Y., Ma, W. & Zhang, H. (2003.) Detecting Web Page Structure for Adaptive Viewing on Small Form Factor Devices. In Proceedings of the ACM Conference on the World Wide Web (WWW'03), Budapest, Hungary, (pp. 225-233).

Freire, J., Kumar, B. & Lieuwen, D. (2001.) WebViews: Accessing Personalized Web Content and Services. In Proceedings of the 10th World Wide Web Conference (WWW '01), Hong Kong.

Hand, T. (1997.) A Proposal for Task-based Evaluation of Text Summarization Systems. Proceedings of the Association for Computational Linguistics / European Association for Computational Linguistics Summarization Workshop, Madrid, Spain.

Hiltz. S. R. & Turoff, M. (1985.) Structuring Computer-Mediated Communication Systems to Avoid Information Overload. Communications of the Association for Computing Machinery, 28(7), (pp. 680-689).

Mani, I. (2001.) Recent Developments in Text Summarization. In Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM'01), Atlanta, Georgia, (pp. 529-531).

McKeown, K., Passonneau, R. J., Elson, D. K., Nenkova, A. & Hirschberg, J. (2005.) Do Summaries Help? A Task-Based Evaluation of Multi-Document Summarization. In Proceedings of the 28th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil.

Mendelsohn, T. (2001.) Content at Point-of-Care: In the Palm of a Doctor's Hand. In Proceedings of Online Information, London, (pp. 169-174).

Milic-Frayling, N. & Summerer, R. (2002.) SmartView: Flexible Viewing of Web Page Contents. In Proceedings of the 11th World Wide Web Conference (WWW '02).

Mitchell, C. C. & West, M.D. (1996.) The News Formula: A Concise Guide to News Writing and Reporting. St. Martin's Press, New York.

Morris, A. H., Kasper, G. M. & Adams, D. A. (1992.) The Effects and Limitations of Automatic Text Condensing on Reading Comprehension Performance. Information Systems Research, 3(1), (pp. 17-35).

Neter, J., Wasserman, W. & Kutnet, M. H. (1990.) Applied Linear Statistical Models, 3rd Edition. Irwin Publishers.

Over, P. & Yen, J. (2003.) Intrinsic Evaluation of Generic News Text Summarization Systems. In Proceedings of the Human Language Technology Conference Workshop on Text Summarization (DUC 2003), Edmonton, Canada.

Radev, D. R., Hovy, E. & McKeown, K. (2002.) Introduction to the Special Issue on Summarization. Computational Linguistics, 28(4).

Radev, D. R., Allison, T., Blair-Goldensohn, S., Blitzer, J., Çelebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., Otterbacher, J., Qi, H., Saggion, H., Teufel, S., Topper, M., Winkel, A., & Zhang, Z. (2004.) MEAD - A Platform for Multidocument Multilingual Text Summarization. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 04), Lisbon, Portugal.

Radev, D. R., Jing, H., Stys, M., & Tam, D. (2004a.) Centroid-based Summarization of Multiple Documents. Information Processing and Management, 40, (pp. 919-938).

Salton, G. & Buckley, C. (1988.) Term-weighting Approaches in Automatic Text Retrieval. Information Processing and Management, 24(5), (pp. 513-523).

Seki, Y., Eguchi, K. & Kando, N. (2004.) Compact Summarization for Mobile Phones. In Crestani, F., Dunlop, M. & Mizzaro, S. (Eds.) Mobile and Ubiquitous Information Access, Lecture Notes in Computer Science Vol. 2954, Springer-Verlag, (pp. 172-186).

Sweeney, S. O., Crestani, F. & Tombros, A. (2002.) Mobile Delivery of News Using Hierarchical Query-Biased Summaries. In Proceedings of the Association for Computing Machinery Symposium on Applied Computing (ACM SAC 2002), Madrid, Spain.

Trevor, J., Hilbert, D. M.,Schilit, B. N. & Koh, T. K. (2001.) From Desktop to Phonetop: A UI for Web Interaction on Very Small Devices. In Proceedings of the 14th Annual Association for Computing Machinery Symposium on User Interface Software and Technology, Orlando, Florida.

Turpin, A. & F. Scholer. User Performance versus Precision Measures for Simple Search Tasks. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’06), Seattle, (pp. 11-18).

White, R., Ruthven, I., & Jose, J. M. (2001.) Web Document Summarization: A Task-oriented Evaluation. In Proceedings of the 12th International Database and Expert Systems Applications Conference (DEXA 2001), Munich.

Yang, C. C. & Wang, F. L. (2003.) Automatic Summarization of Financial News Delivery on Mobile Devices. In Proceedings of the ACM Conference on the World Wide Web (WWW’03), Budapest, Hungary, (pp. 225-233).

Yang, C. C. & Wang, F. L. (2006.) An Information Delivery System with Automatic Summarization for Mobile Commerce. Decision Support Systems (in press).

Yin, X. & Lee, W. S. (2004.) Using Link Analysis to Improve Layout on Mobile Devices. In Proceedings of the ACM Conference on the World Wide Web (WWW’04), New York, (pp. 338-344).