

# Summary in Context: Searching Versus Browsing

DANIEL M. McDONALD and HSINCHUN CHEN  
The University of Arizona

---

The use of text summaries in information-seeking research has focused on query-based summaries. Extracting content that resembles the query alone, however, ignores the greater context of the document. Such context may be central to the purpose and meaning of the document. We developed a generic, a query-based, and a hybrid summarizer, each with differing amounts of document context. The generic summarizer used a blend of discourse information and information obtained through traditional surface-level analysis. The query-based summarizer used only query-term information, and the hybrid summarizer used some discourse information along with query-term information. The validity of the generic summarizer was shown through an intrinsic evaluation using a well-established corpus of human-generated summaries. All three summarizers were then compared in an information-seeking experiment involving 297 subjects. Results from the information-seeking experiment showed that the generic summaries outperformed all others in the browse tasks, while the query-based and hybrid summaries outperformed the generic summary in the search tasks. Thus, the document context of generic summaries helped users browse, while such context was not helpful in search tasks. Such results are interesting given that generic summaries have not been studied in search tasks and that the majority of Internet search engines rely solely on query-based summaries.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process, selection*

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Summarization, search, browse, generic summaries, information seeking, indicative summaries, text processing, natural language processing

---

## 1. INTRODUCTION

Today's Web search engines use text summaries to help users make relevance decisions. Most summaries, such as those used by Google, are based on the query

---

We thank the NSF Digital Library Initiative-2 for funding "High-Performance Digital Library Systems: From Information Retrieval to Knowledge Management," IIS-9817473, April 1999–March 2002.

Authors' addresses: Department of Management Information Systems, University of Arizona, Room 430, McClelland Hall, 1130 E. Helen Street, Tucson, AZ 85721; email: {dmm,hchen@eller.arizona.edu}.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or direct commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 1515 Broadway, New York, NY 10036 USA, fax: +1 (212) 869-0481, or [permissions@acm.org](mailto:permissions@acm.org).

© 2006 ACM 1046-8188/06/0100-0111 \$5.00

terms from the user's search. Basing a summary on the query terms used in the search makes the summary more focused to the user's query and less so to the context of the page. Page-level context information, however, may be helpful to the user when making relevance decisions. Context has been shown to be more important when a user is not as familiar with a search topic and their ideas are still fuzzy or when a user seeks background information [Marchionini and Shneiderman 1988; Kuhlthau 1991; Carmel et al. 1992]. Despite these well-researched needs for context, text summaries provided by Web search engines remain consistently query based at the expense of page-level context.

Web searchers have been shown to distinguish between searching and browsing tasks [Hearst 2002]. Given such user ability, search engines could provide various types of summaries based on the user's need given the information-seeking task being performed. To the best of our knowledge, generic summaries have not been extrinsically evaluated given an information-seeking task. In the Document Understanding Conferences (DUCs), which started in 2001 [Over and Yen 2004], generic summaries have been evaluated intrinsically, being compared to human-generated summaries. In the TIPSTER Text Summarization Evaluation Conference [SUMMAC; Mani et al. 1998], the first large-scale summarization evaluation conference held in May 1998, only query-based summaries were evaluated in a search scenario. Generic summaries were given an extrinsic categorization task. The value of page-level context, as provided by a generic summary, in searching has thus not been explored. In this article, we start by reviewing the relevant summarization and information-seeking literature. We then present the design and intrinsic evaluation of a generic summarizer. The generic summarizer is then used along with query-based and hybrid summarizers in an information-seeking experiment involving 297 subjects. The subjects completed search and browse tasks, each using a different type of summary. The article ends with results and conclusions from the experiment.

## 2. LITERATURE REVIEW

A common way to separate research in automatic text summarization is by the focus or scope of the summary being produced [Firmin and Chrzanowski 1999]. Summaries that ignore the summary user, focusing more on the essence of the document, are "generic" summaries. Summarizers that focus the summary on a topic or user are "query-based" summaries. This distinction is important as it often determines what summarization techniques to apply, what type of document information to utilize, and what type of evaluation method to use. Below, we review the literature on summarization techniques for each type of summary. We then review evaluation methodologies for each type of summary, focusing specifically on information retrieval tasks. We end the literature review by discussing the different types of information-seeking tasks and how document context has been shown to be useful in browsing tasks.

### 2.1 Generic Summaries

Because generic summaries are not based on particular topics, techniques for creating generic summaries strive to include the most important sentences from

a document based on indicators from the document itself. Techniques for identifying sentences fall into two general categories. The first category of generic summarization techniques creates groups of text that represent topic areas. Such textual areas usually span sentences, not relying on sentence boundaries to determine their size. Various types of discourse analysis and information extraction techniques create the textual groupings. The second technique groups text based on sentence boundary markers. This sentence-selection heuristic approach selects sentences for a summary based on preestablished features or characteristics of a sentence. These heuristics are often calculated using surface-level analysis and information extraction techniques. We will first review techniques that place document text into topic groups, followed by those that use sentence-level techniques. We then review how information extraction techniques contribute to both approaches.

Discourse analysis refers to utilizing document meaning and structure that come from text longer than one sentence [Liddy 1998]. If topic areas can be identified, then representative sentences from various topics can be included in the summary. Topic areas can be identified by analyzing document content or document structure. Content analysis focuses on the keywords in the document and the identification of coreferents. Structural analysis, also called *coherence analysis*, uses regularities in document structure to group sentences together.

**2.1.1 Topic Finding Using Document Content.** Cohesion analysis, a type of discourse analysis, refers to measuring the connectivity sentences have to each other or the reliance they have on one another. In text summarization, word cooccurrence information and coreference analysis approximate the cohesion of a document. In SUMMAC, nine of the sixteen participants utilized cooccurrence information, including researchers from TextWise, the University of Massachusetts, and Cornell University [Mani et al. 1998]. Sentences with large numbers of cooccurring words have a greater likelihood of sharing a topic. Lexical chaining [Barzilay and Elhadad 1999] is similar to cooccurrence analysis. Words between sentences are chained together based on their shared semantics. In coreference resolution, sentences are linked together through anaphora or other types of reference resolution. The existence of pronouns from different sentences having the same referent shows cohesion between the sentences. An example of such approach can be found in Boguraev and Kennedy [1997]. In SUMMAC, seven of the sixteen participants utilized coreference techniques to produce summaries. TextWise, General Electric Research, and British Telecom used both coreference and co-occurrence [Mani 1998].

Another technique used for breaking documents into topic groups is text segmentation. TextTiling [Hearst 1997] is an example of a popular text segmentation algorithm. In linear text segmentation, topic groups occur adjacent to one another. Other researchers have improved upon Hearst's algorithm. Kan et al. [1998] considered, among other things, the word classes of terms and semantic clustering of terms using WordNet 1.5 in their segmenting algorithm, SEGMENTER. More recently, Choi [2000] introduced a new algorithm for text segmentation that increased the accuracy of the TextTiling and SEGMENTER algorithms. Additional techniques such as sentence clustering have been used

to group sentences into topics. Nomoto Matsumoto [2001] and Radev et al. [2000] presented topic identification strategies based on clustering. Once topic areas have been identified, summary sentences can be selected from different topics [McDonald and Chen 2002]. Summarization techniques that rely on segmentation usually do not combine traditional surface-level analysis with the topic boundary information.

*2.1.2 Topic Finding Using Document Structure.* A document's structural information is largely obtained using coherence analysis. Coherence analysis, a type of discourse analysis, captures structural information by identifying relationships between sentences or clauses. Marcu [2000] used rhetorical structure theory (RST) trees to represent the relationships between elementary clauses in text. The nucleus nodes on the tree were considered more salient (and thus relevant) than the satellite nodes for a summary. Strazalkowski et al. [1998] utilized the concept of Discourse Macro Structure (DMS) to capitalize on regularities of organization and style in text to choose the best summary sentences. Again, by separating sentences into groups, sentences that cover different topics can be added to the summary.

*2.1.3 Surface-Level Analysis.* The remaining techniques of generic text summarization are the oldest [Luhn 1958] and most commonly used methods. These techniques use surface and entity-level analysis. At this level of analysis various features of sentences are calculated. Feature values can add to or subtract from the weighting of a sentence in a document. Researchers have validated individual features as well as combinations of features that can be identified at this level of analysis. Luhn in 1958 first utilized word-frequency-based rules to identify sentences for summaries [Luhn 1958]. Edmundson [1969] added three rules in addition to word frequencies for selecting sentences to extract, including cue phrases (e.g., significant, impossible, hardly), title and heading words, and sentence location (words starting a paragraph were more heavily weighted). The ideas behind these older approaches are still referenced in modern text extraction research. Teufel and Moens [1999] found the use of cue phrases to be the best individual method. Kupiec et al. [1995] reported that the best mix of extraction features included the position of a sentence within a paragraph, the existence of cue phrases in a sentence, and the sentence length. They found first sentences of paragraphs to be good summary sentences along with sentences containing common cue phrases such as *in summary* and *in conclusion*. Sentences that were simply longer also summarized better. In other research, summary sentences were found to have 90% more proper nouns per sentence [Goldstein et al. 1999]. Identifying the number of "signature words" in each sentence by using the common  $tf \times idf$  calculation was also shown to be a positive contributor to summary quality as well [Aone et al. 1999].

*2.1.4 Information Extraction Techniques.* Information extraction techniques are used in summarization research to enhance topic-finding and surface-level extraction techniques as well as to create task-based summaries. Noun phrasing can be used instead of single words to calculate a sentence's  $tf \times idf$  value, and entity extraction techniques can be used to identify proper

nouns in a sentence. Such techniques enhance the calculation of surface-level measures. Also, as mentioned above, generic summarization techniques have used coreference resolution to chain together sentences of the same topic.

More significant, however, is the role information extraction techniques play in identifying information that can be extracted to support a particular task. In the biomedical domain, for example, genetic interactions can be extracted to assist biologists in constructing gene pathways [McDonald et al. 2004]. The extracted interactions summarize gene pathway information from the document. More general information extraction techniques, including the extraction of template relations, template elements, and scenarios templates, were tested in 1998 at the Message Understanding Conference 7 [MUC-7; Aone et al. 1998]. In general, extracted information must match a predefined structure or template, whether semantic, syntactic, or some combination, in order to be extracted [McDonald et al. 2004].

Task-based summaries differ in several important ways from generic and query-based summaries. First, task-based summaries are *informative summaries* that replace the need to view the entire document. The information extracted (whether facts or events) contains the information sought by the user. On the other hand, generic and query-based summaries are *indicative summaries*, which are created with the intent of helping users make relevance decisions. In addition, information extraction techniques are well suited for multidocument summary creation because they extract uniform information structures. The topic-finding techniques and sentence-selection heuristics we have described above, however, apply mainly to single-document summary creation.

## 2.2 Query-Based Summaries

Query-based summaries focus their analysis at the entity or term level of a document. Words and/or phrases from sections of text are compared to a user's query terms for similarity [Sanderson 1998]. Query-based summaries typically ignore the cohesion or structural similarity between different sections of text, focusing only on a section's query-term similarity. Another approach that utilizes keywords from documents is Maximum Marginal Relevance (MMR) [Carbonell and Goldstein 1998]. MMR views sentences as having redundancy and diversity [Katz 1996]. MMR-based summaries have high redundancy (or relevance) to the query terms and low redundancy to each other. An MMR-based approach strives to present sentences as diverse as possible, but still related to the query.

Keyword-in-context (KWIC) or snippet summaries are query-based summaries that show which query terms appear in a document and the words around those query terms. KWIC summaries used to be rare in information retrieval systems and search engines because of the need to cache the documents locally in order to compare them to query terms [Hearst 1999a, 1996]. However, Google has greatly popularized this summarization method. Some research has shown that text fragments that occur near the beginning of a document and those that contain the largest subset of query terms are the best to include

[Kupiec et al. 1995]. To promote the readability of the summary, the extracted fragments of text are shown in the original document order. The highlighting of the query terms in the summary has been shown to be a useful feature for displaying a KWIC summary [Landauer et al. 1993].

### 2.3 Evaluating Summaries

Generic and query-based summaries are usually evaluated in different ways. Text summary evaluations fall generally into two categories, intrinsic and extrinsic [Mani and Maybury 1999].

**2.3.1 Generic Summary Evaluation.** Generic summaries are most often evaluated intrinsically. Intrinsic evaluations measure the quality of a summary against an ideal. Evaluators compare the summary to human-generated summaries [Kupiec et al. 1995]. The evaluator can scan the summary for the inclusion of key “ideas” from the original text [Brandow et al. 1995] or check for fluency [Minel et al. 1997]. However, problems arise when comparing machine-generated summaries to an “ideal” because there is no single correct summary. Techniques for reducing the subjectivity of the process include creating the “ideal” summary by taking the majority opinion of the experts. Jing et al. [1998] provided a more complete discussion of intrinsic evaluations and their challenges. In the Document Understanding Conferences (DUCs), generic summaries are evaluated intrinsically, being compared to human-generated summaries. In the Summarization Evaluation Conference (SUMMAC), generic summaries were evaluated using an extrinsic categorization task.

**2.3.2 Query-Based Summary Evaluation.** Query-based summaries are most often evaluated with extrinsic tasks. Extrinsic evaluations are task-based evaluations that measure the quality of a summary by testing the user’s ability to complete some other task. For example, users may have to answer questions based on the summary text alone [Morris et al. 1992] or determine the relevance of a document using only the summary [Brandow et al. 1995]. In DUCs, query-based summaries are evaluated using an information-seeking task. A similar information-seeking task was used at the SUMMAC conference. Because selection of summary sentences is based on the query-terms used in a search, information-seeking tasks are an intuitive way to evaluate query-based summaries.

### 2.4 Review of Summaries for Information Seeking

Text summaries have frequently been evaluated using tasks common to information retrieval [Mani et al. 1998]. *Informative summaries*, for example, have been evaluated as document replacements for creating search indexes and producing relevance feedback [Lam-Adesina and Jones 2001]. More common, even, is the evaluation of indicative summaries using an information-seeking task. *Indicative summaries* are not meant to replace the original document, but to provide enough information for the user to judge the relevance of the document [Firmin and Chrzanowski 1999]. In information-seeking tasks, summaries should help users judge document relevance. The type of

information-seeking task being performed, however, affects the relevance decisions [Schamber et al. 1990; Harter 1992]. We will describe two distinct types of tasks in the information retrieval domain, review how subjects have utilized document context in browsing tasks, and discuss the use of document context in information-seeking tools.

**2.4.1 Information-Seeking Tasks.** The process of information seeking involves more than one task. Two general tasks have largely been agreed upon in the information retrieval literature, namely, search and browse. Kuhlthau's [1991] work with high school students showed that students perform different tasks when looking for information. Kuhlthau found that at assignment inception users engaged in more general browsing, with more directed search occurring as topic understanding increased. She found users' thoughts to be general and vague in the beginning of the search process while focused near the end. Other researchers have characterized differences between search and browse in similar ways. Cove and Walsh [1998] presented a three-stage model with 'knowledge of the goal' being the primary factor dictating the information seeking stage. Directed search is performed when the information need is well understood, while general and serendipitous browsing is done when the information need is less clear. Marchionini and Shneiderman [1988] characterized the difference between searching and browsing by the focus of the task. Searching was characterized as more directed and focused, while browsing was described as "an exploratory, information-seeking strategy that depends on serendipity" [Marchionini and Shneiderman 1988, p. 71]. Browse goals had general objectives, while search goals had specific objectives. We end by reviewing the use of page context in browsing scenarios and discussing the lack of context in textual summaries.

**2.4.2 Document Context.** Document context normally contains content not related to the user's query terms, but rather related to the purpose and meaning of the document as a whole, somewhat like the author's perspective. Document context has been shown to be relevant in certain information-seeking tasks. Carmel et al. [1992] used GOMS analysis to study browsing behavior in a hypertext system. Browsing patterns observed were placed into three categories separated in part by how deep the user would delve into any one page. Of the three browsing tasks observed, the task involving deeper open-ended analysis of a document was the most common. In other words, while browsing, users relied on page context to study a document. Analyzing page context appeared to help users affect their own mental context for a topic. Also, in research from natural language processing, Black et al. [1992] noted that humans effectively deal with the ambiguities of natural language sentences by examining the context.

**2.4.3 Single-Document Context Ignored in Browsing Tools.** Web-browsing tools are largely based on two different types of Web mining: Web structure mining and Web content mining [Etzioni 1996]. Web structure mining, also referred to as *citation analysis*, involves using the hyperlink information between documents to show similarity, relevance to queries, or other relationships between documents. Two well-known algorithms that utilize Web structure are

the HITS and Page Rank algorithms [Brin and Page 1998; Kleinberg 1999]. In Web content mining, documents are related to each other or ranked by comparing keywords between documents. The keywords of a single page are all treated equally and compared to keywords from other documents. Tools that group documents together based on the page content include the hierarchical categories of Yahoo and visualization techniques such as the self-organizing map (SOM) [Chen et al. 1996, 1998]. Thus, existing browsing tools mix all the content (both query relevant and context) from each page together to represent the search “landscape.” In such a representation, the document context (or author’s perspective) of a single document contrasted with its query-relevant (searcher’s perspective) content is lost in the aggregation process. To utilize document-level context, an algorithm must separate the context from query-related content.

### 3. RESEARCH QUESTIONS

The main question in our research was whether summaries that include more document context, such as generic summaries, are better suited than query-based summaries in either a browse or directed-search information-seeking task. We explored whether document context as part of a summary helps users make relevance decisions better than adding additional information similar to a user’s query. We first developed and intrinsically evaluated a generic automatic text summarizer, the Arizona Summarizer. Such evaluation was meant to legitimize our techniques for generic summary creation and show validity. Second, we added different variations of query-based summaries and conducted a user study of 297 subjects comparing their performance given search and browse tasks. Such a large user study was conducted because generic summaries have not been evaluated using directed-search tasks before.

#### 3.1 Summarizer Development

We had two research questions regarding the summarizer development. First, we explored how to construct a generic summarizer that balances the goal of drawing sentences from multiple document topics and at the same time containing good summarizing sentences. Systems at SUMMAC most often used cooccurrence and coreference for cohesion analysis. We explored how to use a segmenting algorithm to identify topic areas and then how to choose sentences from within topics once the topic areas are identified. Second, we studied what the resulting performance was of such a summarizing tool. We measured the performance of the generic summarizer intrinsically by benchmarking it against the human-generated summaries of TREC documents compiled by Jing et al. [1998] in their research. We tested whether the performance of the AZ Summarizer makes it generalizable to other systems and thus a good candidate for use in a large user information-seeking study.

#### 3.2 Summary Type and Task Experiment

We explored the usefulness of document context given search and browse tasks. More document context means users see content most related to the

```

sort sentences by weight
while (desiredSumLength is not met and there are unused sentences)
  for (all sentences x )
    if (sentence x not already in summary)
      if (segment of sentence x has the lowest or equally low use)
        set sum_sentence = x
        break out of for loop
      end if
    end if
  end for
  add sum_sentence to the summary
  record sum_sentence as having been used
  increment sum_sentence's segment use
  increment currentSumLength
end while

```

Fig. 1. Pseudocode for extraction algorithm.

document overall and not that just related to the user's query. Generic summaries provide more context than query-based summaries because they include sentences from multiple topic areas or structural areas and do not focus solely on the content most related to the user's query. Other summaries reduce context even more by producing a query-based summary that does not use full sentences, but rather only a small amount of text around the query terms, known as a *snippet summary*. Users are very often uncertain about their choice of query terms, especially during the early stages of the information-seeking process [Kuhlthau 1991]. Showing topics only related to query terms might overbias a summary. In addition, providing content only relevant to the query may exclude information central to the document's meaning or purpose. We explored the question of how the reduction of page-level context impacts user information-seeking performance given search and browse tasks.

#### 4. AZ SUMMARIZER DESIGN

The Arizona (AZ) Summarizer utilizes the boundaries of a document's topic areas along with the common surface heuristics from summarization research. The summarizer uses the TextTiling algorithm to calculate topic area information. Because we also relied on surface-level summarization techniques, the TextTiling algorithm offered adequate performance for linear segmentation. Segmenting algorithms have not included sentence-selection heuristics to select individual sentences beyond that of using term-level calculations [Kan et al. 1998]. The combination of segmentation with surface-level heuristics represents the contribution of our algorithm. The number of topic areas in a document reflects the document's diversity of topics. Because generic summaries are not biased by a particular topic, we implemented an algorithm that extracts equal numbers of sentences from each of the text segments. The pseudocode for the algorithm is shown in Figure 1. Because ideal *compression levels* (ratio of summary length to source length) for a document may change as users' tasks change, flexibility to handle changes in summary length is built into the summarization algorithm. After sorting the sentences by weight and

AZ Summarizer					
Sentence Rank	Sentence No.	Document	Topic Segment	Order Extracted	Summary Order
8	1	-----	1		
1	2	-----	1	1	1
15	3	-----	1		
14	4	-----	2		
12	5	-----	2		
9	6	-----	2		
3	7	-----	2	3	2
6	8	-----	2	5	3
13	9	-----	3		
7	10	-----	3		
10	11	-----	3		
5	12	-----	3		
2	13	-----	3	2	4
11	14	-----	3		
4	15	-----	3	4	5

Fig. 2. Example of the extraction process for a five-sentence summary.

entering the while loop, in the pseudocode, the algorithm extracts a sentence from a topic segment if it has not already been used and if its text segment is the least represented in the summary. The highest-ranking sentence from the least-used segment is always added to the summary first. Two sentences cannot be extracted from the same topic segment until all topic segments have produced a sentence for the summary. This process continues until the summary contains the desired number of sentences or the document contains no unused sentences. Once the algorithm has extracted the desired number of sentences, it positions them in their original document order to promote fluency.

Figure 2 shows an example of how the summarizer would select sentences from a document given that it was segmented into three topic areas. A dashed line represents each sentence in the document (there are 15 in total). The topic segment number to which each sentence belongs is listed on the right. The rank of each sentence as scored by the selection heuristics is listed in the first column (Sentence Rank). The pseudocode in Figure 1 utilizes the sentence rank to sort the array. The algorithm keeps track of the most underutilized segment by using the topic segment information also listed in Figure 2. The Order Extracted column shows what iteration of the pseudocode extracted what sentence. Once sentences are extracted, they are reordered to match their original document order, shown in the Summary Order column.

#### 4.1 Structural Analysis

We used text segmentation to partition the document into multiple topic sections that spanned from one to many sentences. In particular, the TextTiling algorithm was the technique we used for segmentation. The TextTiling algorithm determines where topic boundaries are located. A topic boundary is the point where the document transitions from one topic to the next. The first step in the TextTiling algorithm is to divide the text into token sequences,

removing any words that appear on a stop list. We used a token-sequence length of 20 and the same stop word list used by Hearst [1997] in her implementation. Token sequences were then combined to form blocks. The first block contained the  $(k+1)$ th token sequence plus  $k$  token sequences before it. The second block contained the  $(k+2)$ th token sequence and the  $k$  token sequences after it. The value for  $k$  used in the AZ Summarizer was 10, the same as used by Hearst [1997]. A similarity algorithm compared blocks to adjacent blocks. After each comparison, both blocks advanced one token sequence, where the algorithm made another comparison. The similarity algorithm returned the similarity as a percentage, which was derived from the number of times the same terms appeared in the two blocks being compared. The Jaccard coefficient was used for the similarity equation, which differed slightly from the normalized inner product equation used by Hearst [1997]. The Jaccard coefficient is shown in Equation (1):

$$S_{i,j} = \frac{\sum_{k=1}^L (w_{ik}w_{jk})}{\sum_{k=1}^L w_{ik}^2 + \sum_{k=1}^L w_{jk}^2 - \sum_{k=1}^L w_{ik}w_{jk}}. \quad (1)$$

$S_{i,j}$  is the similarity between the two blocks of grouped token sequences  $i$  and  $j$ . The variable  $w_{ik}$  is thus the total of the occurrences of term  $k$  in block  $i$  and  $w_{jk}$  is the total of the occurrences of term  $k$  in block  $j$ . In the numerator, the occurrence of unique terms ( $k$ ) in both token sequences  $i$  and  $j$  are multiplied together and summed over the set of total unique terms from both blocks,  $L$ . In the denominator of the equation, the occurrence of all words in the two token sequences is squared and totaled and the value of the overlapping words is subtracted out. Thus the similarity is a ratio of shared words to nonshared words. Where the similarity between two token sequences is under a threshold, a segment boundary is created. Once segment boundaries are identified, sentences are assigned to a segment. The segment containing the first word in the sentence is the segment for the entire sentence when segment boundaries appear in the middle sentences.

#### 4.2 Sentence and Entity-Level Analysis

Sentence and entity-level analysis begins with a rule-based sentence boundary detector or sentence splitter. The sentence splitter recognizes 208 common abbreviations and uses rules to find periods that have been placed inside quotation marks or parenthesis. Because abbreviations present a specific problem there are also rules to anticipate abbreviations not in our lexicon. Once sentence boundaries are determined, the order of the sentences is recorded and each sentence is scored based on five sentence-selection heuristics. In selecting which heuristics to utilize, we incorporated as many as possible, knowing some heuristics would be more or less useful in different domains. Equation (2) shows the five heuristics that make up the sentence value ( $SV$ ) for sentence  $k$ ,  $SV_k$ :

$$SV_k = a_1 \times S_{cp}(k) + a_2 \times S_{pn}(k) + a_3 \times S_{sw}(k) + a_4 \times S_{sp}(k) + a_5 \times S_{sl}(k). \quad (2)$$

Table I. Impact of Sentence Selection Heuristics

Sentence-Selection Heuristics	Points Allotted
1 Cue phrase	40 points
Proper nouns make up 27% of sentence words	34 points
$tf \times idf$ normalized for the length of the sentence	20–50 points a common range
Sentence begins document	30 points
Sentence begins paragraph	20 points
Sentence length of 358 characters	40 points

$S_{cp}(k)$  is the cue phrase value for sentence  $k$ ,  $S_{pn}(k)$  is the proper noun value for sentence  $k$ ,  $S_{sw}(k)$  is the signature word value for sentence  $k$ ,  $S_{sp}(k)$  is the sentence position value for sentence  $k$ , and  $S_{sl}(k)$  is the sentence length value for sentence  $k$ . The total of each heuristic is multiplied by a weighting factor ( $a_1 \cdots a_5$ ) to normalize the impact of any one score. Table I shows the points allotted to a sentence given the corresponding heuristic values.

We manually adjusted the weights through a process of summarizing a training set of information technology articles and adjusting the weights after assessing each document to balance the impact of any one heuristic. The cue phrase heuristic was slightly weighted above the others. The other four heuristics were adjusted to have somewhat similar impact on the ranking of a sentence. Because the value of the weights is dependent on the domain of articles being summarized, we wanted to avoid overcommitting to a particular domain. Also, it has been reported that the weighting of different summarization features does not significantly affect the average precision [Lam-Adesina and Jones 2001]. Details of each heuristic are described below.

**4.2.1 Cue Phrases.** Each sentence was checked for the existence of 10 different cue phrases (e.g., *in summary*, *in conclusion*, *in short*, *therefore*). Cue phrases indicate where an author may intend to summarize an idea. This heuristic has been weighted heavily because cue phrases are rare and have been shown to be good indicators of summary sentences [Teufel and Moens 1999]. If a sentence has a cue phrase, the sentence value should be high regardless of the totals from the other four heuristics. The calculation of the cue phrase value for sentence  $k$  is shown in Equation (3):

$$S_{cp}(k) = \sum_{t_i \in s_k} w_{cue}(t_i). \quad (3)$$

The value  $w_{cue}(t_i)$  is the weight of the term  $t_i$  found in the cue phrase dictionary. Values in the dictionary range from  $-1$  to  $1$ . Currently, all cue phrases in the dictionary have a weight of  $1$ . All phrases not found in the dictionary are given a weight of zero. We may include phrases with a negative cue weight value in the future. A sentence's cue phrase value is the sum of the cue weights of the sentence's phrases,  $s_k$ . The cue phrase value is then multiplied by a weight  $a_1$  to normalize its impact on the value of the sentence.

**4.2.2 Proper Nouns.** An indicative summary is meant to provide enough information for a user to decide whether to read the original document in its entirety. Proper names and places impact such relevance decisions. The

importance of proper nouns, however, varies between domains. To obtain a rough estimate of proper nouns in a sentence, we count capitalized words, not including the opening word in the sentence. While capitalized words do not always equate to proper names and places in our formatted information technology (IT) news domain, the correspondence is acceptable. A full entity-extraction system could be used to replace the reliance on word capitalization and even differentiate between the types of proper nouns in each sentence. The calculation for the scoring of proper nouns found in a sentence is shown in Equation (4):

$$S_{pn}(k) = \frac{\sum_{t_i \in s_k} p_{noun}(t_i)}{|s_k|}. \quad (4)$$

The value  $p_{noun}(t_i)$  is the proper noun value of term  $t_i$ . Proper noun values are either 1 if the proper noun is recognized or zero otherwise. The proper noun value for sentence  $k$  is the sum of the proper noun values given to each term,  $t_i$ , in the sentence,  $s_k$ . The sum of the proper noun values is divided by the total number of words in the sentence,  $|s_k|$ . Thus, shorter sentences are not penalized for having fewer proper nouns than longer sentences.

**4.2.3 Signature Words.** The formula  $tf \times idf$  or term frequency multiplied by inverse document frequency measures how common the words in a sentence are relative to the entire document. Signature words are words that are common to a sentence/document, but not to a document/collection. Sentences with more signature words are scored higher. The formula to calculate scores of average signature words per sentence is shown in Equation (5), where  $w_{ik}$  is the  $tf \times idf$  score for term  $t_i$  in sentence  $k$ :

$$S_{sw}(k) = \frac{\sum_{t_i \in s_k} w_{ik}}{|s_k|}. \quad (5)$$

The  $tf \times idf$  score for each term  $t_i$  from sentence  $s_k$  is summed and divided by the total number of words in the sentence  $|s_k|$  to create an average. The calculation of  $w_{ik}$  is shown in Equation (6), the  $tf \times idf$  score for term  $i$  in sentence  $k$ :

$$w_{ik} = tf_{ik} \times \log_2 \left( \frac{N}{n} \right). \quad (6)$$

In the formula,  $tf_{ik}$  is the term frequency of term  $i$  in sentence  $k$ .  $N$  is the total number of sentences in the document, and  $n$  is the number of sentences in the document, which contain  $t_i$  (term  $i$ ). Before term frequencies and document frequencies are totaled, each word is made lowercase and stemmed using the Porter stemmer. The Porter stemmer is one of the most widely used stemming algorithms [Jurafsky and Martin 2000] and can be thought of as a lexicon-free stemmer because it uses cascaded rewrite rules that can be run very quickly and do not require the use of a lexicon. Stemming is performed so that words with the same stem but different affixes may be treated as the same word when calculating the frequency of a particular term.

4.2.4 *Sentence Position in a Paragraph.* As the sentences are extracted from the original document, new lines and carriage returns signal the beginning of new paragraphs. The beginning sentence of a document and the beginning sentence of each paragraph are scored higher due to their greater summarizing potential. The calculation of the sentence position score for sentence  $k$  is shown in Equation (7):

$$S_{sp}(k) = S_{pos}(P_k). \quad (7)$$

$P_k$  is the position of sentence  $k$  in the document, and  $S_{pos}(P_k)$  is the value given to that position. Position values range between zero and 1. Currently, sentences that start a document or a paragraph are given the value of 1. All others are given the value of zero.

4.2.5 *Sentence Length.* The length of a sentence can provide clues as to its usefulness in a summary [Kupiec et al. 1995; Aone et al. 1999]. The sentence length calculation is shown in Equation (8):

$$S_{sl}(k) = S_{len}(L_k). \quad (8)$$

$L_k$  is the number of characters in sentence  $k$ , and  $S_{len}(L_k)$  is the value given to the length of sentence  $k$ . Before increasing sentence score based on sentence length, we tried to achieve the same effect by not averaging  $tf \times idf$  scores over the number of words in the sentence. Longer sentences received higher scores by virtue of their greater number of signature terms. However, this approach disproportionately weighted long sentences. Sentence length became its own heuristic to remedy this problem.

## 5. AZ SUMMARIZER EXTENSIONS

In addition to the generic AZ Summarizer, we developed a full-sentence, hybrid summarizer and a snippet, query-based summarizer. The hybrid and the query-based summaries were created to study the impact of varying amounts of discourse analysis and document context within summaries that were still based on the query terms. The performance of all three summaries could then be compared in different information-seeking tasks.

### 5.1 AZ Full-Sentence, Hybrid Summary

In the full-sentence, hybrid summary, the five-sentence-selection heuristics were replaced with a single  $tf \times idf$  score of the query terms in the document. Each sentence was treated as a document, with all the sentences together composing the collection. The  $tf \times idf$  equation used was in Equation (6). Term frequency,  $tf_{ik}$ , was the occurrence of query term  $i$  in sentence  $k$ .  $N$  was the total number of sentences in the document, and  $n$  was the number of sentences with the query term. The  $tf \times idf$  score was summed over all the query terms. The topic information provided by the TextTiling algorithm was still utilized in this summary. Summary sentences were still chosen from different topic areas, but sentence ranking was based solely on the similarity of the sentence to the query (the  $tf \times idf$  score). Because of its query focus, this summary provides less

document-level context than the generic summary, but more than the snippet summary because it uses full sentences coming from different topic areas as determined by discourse analysis.

## 5.2 AZ Snippet, Query-Based Summary

The snippet, query-based summary did not use any structural information obtained through TextTiling, nor did it include full sentences. The snippet summary, also referred to as *keyword in context* (KWIC) [Hearst 1999a,1999b] in other research, was created by locating the query terms in the document and including the adjacent 55 characters of text on each side of the query term. To match the length of the full-sentence summaries, three total instances of query terms and their snippets were extracted from the document. Depending on how often the different query terms appeared in the document, the summary would include snippets from all the query terms or three different snippets using the same query term. The snippets were concatenated together, separated by three dots. The query terms in the snippets were bolded as is common with Internet search engines. This type of summary was included in the experiment because it utilized no discourse analysis and contained less document context than the full-sentence summaries. In addition, query-based summaries are commonly used on the Internet by search engines such as Google. An example of each of the summaries is shown in Appendix A.

## 6. RESEARCH HYPOTHESES

Our research hypotheses fell into two major categories. The first involved the evaluation of the generic AZ Summarizer using an intrinsic study. The second involved the extrinsic evaluation of four different types of summaries using two types of information-seeking tasks.

### 6.1 Generic AZ Summarizer Performance

We had one formal hypothesis to test regarding the performance of the AZ Summarizer design:

- H1: The AZ Summarizer will perform the same or better than at least two published summarization systems evaluated on the TREC dataset at the 20% compression level.

### 6.2 Summaries of Varying Page-Level Context in Information-Seeking Tasks

We tested five formal hypotheses regarding the value of page-level context found in text summaries given different information-seeking tasks:

- H2: The generic summary, which has more page-level context than the hybrid and query-based summaries, will perform better than the hybrid and query-based summaries in browse tasks.
- H3: The hybrid summary and query-based summary will perform better than the generic summary in search tasks.

- H4: Full-sentence, hybrid summaries have more page-level context and thus will outperform query-based snippet summaries in browse tasks.
- H5: Snippet, query-based summaries will outperform full-sentence, hybrid summaries in search tasks.
- H6: The generic summary will outperform all other summaries overall.

## 7. EXPERIMENTAL DESIGN

In order to test the six hypotheses above, we designed two experiments.

### 7.1 AZ Summarizer Experiment: Intrinsic Evaluation

Using human-generated summaries as a gold standard is termed *intrinsic evaluation* and is the most common summarization evaluation technique. We obtained a well-established corpus of human-generated summaries of TREC documents used in previous research [Jing et al. 1998; Marcu 2000]. In this experiment, the computer-selected sentences were compared to the human-selected sentences for recall and precision at two different levels of compression. Compression is the ratio of summary length to the length of the source document. Previous research has recognized the shortcomings of this approach, explaining that there is no single correct summary [Jing et al. 1998]. To minimize such shortcomings, Jing et al. had five subjects create summaries of 40 news articles from the TREC collection. Each subject created a summary at 10% and 20% compression levels, resulting in 400 total summaries. A summary of 10% compression is 10% the length of the original document. Sentences with the highest percent agreement were used to create the ideal summary for the 40 articles. Percent agreement is the ratio of observed agreements with the majority opinion to the possible agreements with the majority opinion. Refer to Jing et al. [1998] for the details of the corpus creation. Summary length has been shown to impact the recall and precision performance of a summarizer. To control for variations of the 10 and 20% compression calculations, we added the same number of sentences to a summary that had been added by human subjects. In other words, the number of relevant sentences in the summary was also the same as the number of retrieved sentences. As a result, our precision and recall numbers were identical. If the human-generated summary had six sentences, so did ours. If four of ours were relevant, both precision and recall were 67 percent (4/6).

Given the corpus of human generated sentences, we calculated recall and precision for three different scenarios. First, we compared the AZ Summarizer sentences to the consensus of ideal sentences. Second, we compared the AZ Summarizer sentences to any of the human-generated summary sentences. Third, we compared sentences randomly extracted to the sentences from the consensus ideal summaries. Because lead sentences in the news domain tend to be very good summarizing sentences, we also compared the lead sentences from the documents to the consensus ideal summary sentences.

Next, we compared the intrinsic performance of the AZ Summarizer to other summarizers that have used the same data to calculate recall and precision. The comparison was made to determine support of our hypothesis that the AZ

Describe the open source movement. (Open ended)  
What is XML? (Open ended)  
What is the name of Novell's annual user conference? (Search)  
What is Bill Joy's position at Sun Microsystems? (Search)

Fig. 3. Tasks for summary experiment.

Summarizer performs the same or better than at least two published summarization systems evaluated on the TREC dataset at the 20% compression level and is therefore generalizable enough to be used in an information-seeking experiment.

## 7.2 Using Greater Page-Level Context in Information-Seeking Tasks

In order to test our hypotheses on information-seeking using summaries, we utilized a collection of approximately 300,000 business information technology documents. We obtained the collection through metasearching various news magazine and editorial Web sites [Marshall et al. 2004]. We tested the effectiveness of four different types of summaries in identifying material relevant to two search tasks and two browse tasks. Such a task-oriented approach is an accepted methodology for evaluating information retrieval systems [Hersh et al. 1996]. Search tasks were narrow, requiring users to find documents with specific facts of one to two words. Browse tasks were broad and open ended, which required users to understand more of the topic in order to choose relevant documents [Marchionini and Shneiderman 1988]. The tasks were patterned after the TREC Ad Hoc Retrieval Tasks, but tailored to the information technology domain. The four tasks are listed in Figure 3.

Six tasks in total were considered for the experiment, two browse and four search tasks. Two query terms were used to represent each task, and all the documents in the collection with the query terms were retrieved. The total number of documents retrieved varied between 20 and 40. Two search tasks were eliminated before the pilot because there were not enough irrelevant documents within the retrieved set. We reviewed the list of document results and selected 12 documents that had varying levels of relevance. At least three documents for each task were highly relevant. We attempted to include at least seven documents that were less relevant, which was difficult in some cases because all the documents contained the two query terms and were part of a highly focused collection that had been constructed through a metasearch collection-building methodology [Marshall et al. 2004].

Users were required to select only the three most relevant documents from the 12 listed. Setting the number of documents to select differs from traditional recall/precision experiments, yet allows subjects to compare the relevance of documents against each other as opposed to a relevance standard that may vary more between subjects. The library expert read the full text of all the documents and selected the three most relevant documents of the 12 for each task. In one task, the expert found five documents to be equally relevant, so all five were considered correct for all conditions in the experiment. Two small pilot studies were run to test the reliability of the Web-based experimental platform,

the clarity of the pre- and postquestionnaires, and the understandability of the tasks and experimental instructions. One browse task was altered after the pilot to be more open ended.

A total of 297 subjects participated in the experiment. The subjects participated in the experiment as part of an introductory course in management information systems. Participation in this experiment was a required part of the course. After logging onto the system, the user was presented with specific instructions for the experiment and then an eight-question prequestionnaire. After the questionnaire, the user was presented with the first of the four tasks. Each of the four tasks was assigned to each subject in random order. Under each task, a list of 12 results was presented in random order. For each of the four different tasks, the results listing used different types of summaries. An example of each of the four types of summaries is included in the Appendix. Each task-summary combination was also randomly assigned to the different subjects. Four tasks were chosen to match the four types of summaries. Using the same number of tasks as summaries was done to balance the impact of any one person on a particular task-summary combination. Every subject saw the same four tasks and used all four summaries. All tasks appeared with the four different types of summaries an equal number of times. Each summary type was selected a total of 891 times (see Table III). A postquestionnaire followed the completion of the four information-seeking tasks.

Subjects were instructed to identify and mark the top three documents (of the 12 results) that provided the most relevant information to complete the task. The correct response to each task was therefore not the answer to the question, but rather to select the top three documents that would be most relevant to answering the question. Of course, if answers to the task did appear in the summary, this was strong evidence on which to base relevance decisions. Subject's relevance decisions were based on the summaries alone. The time spent on each task given the summary type was kept. To serve as the gold standard, an expert identified the top three relevant documents for each of the four tasks based on the entire document without referring to the summaries. The expert had previously conducted such relevance evaluations of documents from medical research and from general Internet searching. The expert had a Masters degree in Library and Information Science and also a Bachelor of Fine Arts and a Master of Arts in Art History. Her library experience also included being a faculty member at the University of Arizona and Clemson University Libraries. Results from the experiment were tallied by totaling the correctly identified documents according to the type of summary used in the decision.

In order to control for different abilities in query formation, two query terms for each task were predetermined and the resulting 12 documents were retrieved. All summaries were generated and stored before the experiment to normalize system response time. The hybrid summary and the query-based summary used the predetermined query terms to generate the summaries.

The first type of summary was a two-sentence generic summary created by the AZ generic summarizer. The generic summary also included the top five

keywords from the page and the number of sentences on the page. The second summary was a two-sentence hybrid summary created by the AZ full-sentence, hybrid summarizer. In addition, this full-sentence, hybrid summary contained how many times the query terms appeared in the document and which of the query terms were found. The third summary was a three- to five-line snippet query-based summary created by the AZ query-based summarizer. The length of the snippet summary was calculated to be as close to the length of the full-sentence summaries as possible. The fourth summary was obtained from the IT news sites that had published the documents. We refer to the fourth summary as the *original summary*, and we included it to serve as a baseline for the performance of the AZ summarizers. The “original” summaries were created using different techniques. Summaries from IDG, Infoworld, and sites powered by Google used snippet query-based summaries. Summaries from C|Net used the lead sentences from the articles, and summaries from Computerworld and PCWorld appeared to be human-generated summaries. The length of the original summaries was usually shorter than the three summaries created by the AZ Summarizers. All summaries also included the title of the Web page and the date of the article, when available. The query terms found in the summaries were all bolded, with the exception of the original summary.

Of the 297 subjects that participated in the experiment, 157 were male and 140 were female. The students had majors somewhat evenly distributed between Finance, Marketing, Management Information Systems, Management, and Accounting. Fifty-six percent of the subjects used mainly Google in their daily searching, followed by 31% for Yahoo, and 3% for MSN. Seventy-seven percent of the subjects were native English speakers, while the other 23 percent had a different native language (20 languages were represented in total).

## 8. EXPERIMENTAL RESULTS

The results and discussion of the intrinsic experiment along with results and discussion of our information-seeking experiment with summaries of differing page-level context now follow.

### 8.1 AZ Summarizer Experimental Results from Intrinsic Experiment

According to our experimental design, we evaluated intrinsically how the AZ generic summarizer performed using human-generated summaries as the gold standard. In this experiment, we fed the automatic summarization routine the number of sentences allowed in each summary to meet the compression requirements. As a result, the percentages for recall and precision were identical (described in Section 7.1). The results of the experiment are summarized in Table II. First, we compared the AZ generic summarizer to the consensus of human-selected sentences. We achieved 50% precision and recall at the 20% compression level and 47% precision and recall at the 10% compression level.

Next, we compared the AZ Summarizer to the most similar subset of human-selected sentences. This gold standard was therefore not a consensus of human-selected sentences, but rather the set of human-selected summaries that best matched those of the AZ Summarizer. Performance went up to 73%

Table II. Performance of Automatically Generated Summaries

Technique	Compression	Precision / Recall
AZ Summarizer compared to consensus ideal summaries	20%	50%
	10%	47%
AZ Summarizer compared to all human-generated sentences	20%	73%
	10%	58%
Random sentences compared to consensus ideal summaries	20%	20%
	10%	5%

Table III. Summarizer Performance on TREC Corpus

Technique	Compression	Precision/Recall
AZ Summarizer	20%	50%/50%
	10%	47%/47%
Marcu's Summarizer [Marcu 2000]	20%	50%/47%
	10%	N/A
Jing et al. Summarizer A [Jing et al. 1998]	20%	32%/39%
	10%	33%/37%
Jing et al. Summarizer B [Jing et al. 1998]	20%	47%/64%
	10%	62%/67%
Jing et al. Summarizer C [Jing et al. 1998]	20%	36%/55%
	10%	46%/64%

precision and recall and 58% precision and recall for the 20% and 10% compression levels, respectively. Therefore, 73% of the sentences selected by the AZ generic summarizer at the 20% compression level were also selected by at least one of the five human summarizers. Finally, a summarizer that randomly selected sentences for extraction achieved poor performance, with a 20% precision and recall total at the 20% compression level and a 5% precision and recall total at the 10% compression level. Also, taking the lead sentences to generate a summary proved quite effective because the TREC corpus used was made up of news documents.<sup>1</sup>

Next, we compared the performance of the AZ Summarizer to that of four other summarizers that were evaluated on the same TREC data set [Jing et al. 1998; Marcu 2000]. While the significance of head-to-head comparisons has been questioned [Jing et al. 1998], we present this comparison only to support our hypothesis that the AZ generic summarizer performs the same or better than at least two published summarization systems at the 20% compression level. Table III summarizes the results of the comparison. At the 20% compression level, the AZ generic summarizer matched the best-reported precision measure of 50%. Recall performance of the AZ Summarizer was ranked right in the middle compared to the recall numbers reported by the four other summarizers, confirming our first hypothesis.

## 8.2 Discussion of Intrinsic Summarization Experiment

The fact that 73% of the sentences extracted by the AZ Summarizer were also extracted by at least one human is promising given the difficulty of

<sup>1</sup>Lead sentence summaries achieved 58% precision and recall at the .20 compression level.

Table IV. Documents Selected by Subjects as Relevant to Browse and Search Tasks

	Browse Tasks				
	Generic	Hybrid	Original	Query based	Browse
Relevant	227 (51%)	144 (33%)	154 (34%)	127 (29%)	652 (37%)
Not relevant	220 (49%)	294 (67%)	305 (66%)	311 (71%)	1130 (63%)
Total	447	438	459	438	1782
	Search Tasks				
	Generic	Hybrid	Original	Query based	Search
Relevant	118 (27%)	145 (32%)	185 (43%)	163 (36%)	611 (34%)
Not relevant	326 (73%)	308 (68%)	247 (57%)	290 (64%)	1,171 (66%)
Total	444	453	432	453	1782
Relevant total	345 (39%)	289 (32%)	339 (38%)	290 (33%)	1263 (35%)
Total selected	891	891	891	891	3564

reproducing human summaries. In addition, the AZ generic summarizer easily outperformed a random summarizer in recall and precision. The AZ generic summarizer performed better at the 20% compression level compared to the 10% level in part because the heavy reliance on lead sentences in the human summaries was lessened as the length of the summary increased. The AZ generic summarizer sought a diverse summary by adding sentences from different topic areas. Such a strategy is less likely to produce summaries, with sentences appearing next to each other as occurs with lead-sentence summaries. Based on the comparable performance of the AZ generic summarizer compared to other summarization tools reported, we can confirm H1. The AZ generic summarizer is generalizable and well suited for an experiment studying the information-seeking benefits of different types of summaries. In addition, the lower concentration of news articles in our IT collection compared to the TREC collection limits the bias toward lead sentences, which may have helped our summarizer perform better.

### 8.3 Experimental Results for Information Seeking with Summaries

We tallied the number of relevant documents selected by summary type and by information-seeking task, whether browse or search. The total numbers are reported in Table IV. A grand total of 3564 summaries were selected by the subjects in the experiment as being relevant (three documents per task). From that total, an expert confirmed 1263, or 35%, of the documents were actually among the top three relevant documents for the task. By separating browse and search tasks and the four different types of summaries, we measured the tendency to make better relevance decisions given a certain type of information seeking scenario and summary.

The results showed that, given the browse task, the generic summary guided users to relevant documents 51% of the time compared to 29% for the snippet summary. Given a search task, the original summary successfully guided users 43% of the time compared to 27% for the generic summary. When ignoring the type of task, the subjects scored 39% accuracy with the generic summary, while achieving 32% accuracy with the full-sentence, hybrid summary. Overall, users were slightly more successful using the summaries in browse tasks (37% accuracy) than in search tasks (34% accuracy).

Table V. Summary Type Results Given a Browse Task

Browse Task	$t$ Stat	$p$ Value	Result
Generic is better than the hybrid summary	7.747874	0.00000	H2 confirmed
Generic is better than original summary overall	7.455501	0.00000	H6 pending
Generic is better than query-based snippet summary	9.502109	0.00000	H2 confirmed

Table VI. Summary Type Results Given a Search Task

Search Task	$t$ Stat	$p$ Value	Result
Query-based snippet is better than generic summary	4.11075	0.0001	H3 confirmed
Hybrid full-sentence is better than generic summary	2.804726	0.00570	H3 confirmed

**8.3.1 Summaries with Browse Tasks.** We conducted a two-tailed  $t$  test comparing the mean totals of relevant documents in the browse task by summary type to see if the totals could have come from the same distribution. The results from the statistical analysis are listed in Table V. The generic summary produced significantly more relevant documents than the hybrid, full-sentence summary ( $p$  value = .0000), the query-based snippet summaries ( $p$  value = .0000), and the original summaries from our content providers ( $p$  value = .0000). These conclusions were also confirmed using analysis of variance (ANOVA), which showed the between-group variation having a significance less than .000 (with an  $F$  score of 20.66 and 3 degrees of freedom).

Table IV contains details of the page totals. Given this statistical evidence, we confirmed our second hypothesis that generic summaries with more page-level context more effectively lead users to relevant documents than query-based and hybrid summaries in browsing tasks. This finding highlights the usefulness of page-level context for browsing tasks. Often, such page-level context is ignored in favor of query-based summaries in Internet search engines. Interesting as well is that the generic summaries outperformed the summaries actually used by content providers in the browsing tasks.

**8.3.2 Summaries with Search Tasks.** Next, we conducted a statistical analysis using a two-tailed  $t$  test comparing the mean performance of the different summary types in the “search” task. The results showed that users were able to identify significantly more relevant documents given the hybrid, full-sentence summary ( $p$  value = .0001) and query-based, snippet summaries ( $p$  value = .0057) than they were given a generic summary. The statistical analysis is summarized in Table VI. These conclusions were also confirmed using analysis of variance (ANOVA), which showed between-group variation having significance less than .000 (with an  $F$  score of 13.44 and 3 degrees of freedom). Our third hypothesis that both full-sentence, hybrid summaries and snippet, query-based summaries would outperform generic summaries was confirmed by the results. Thus summaries that are based on the queries alone are more useful given focused search tasks.

**8.3.3 Full-Sentence Hybrid Versus Snippet Query-Based Summaries.** Our fourth hypothesis was that full-sentence, hybrid summaries would outperform snippet, query-based summaries in a browse task. We reasoned that hybrid

Table VII. Full-Sentence Versus Snippet Summary

Browse	$t$ Stat	$p$ Value	Result
Full-sentence hybrid is better than snippet query-based summary	1.840058	0.0678	H4: marginal confirmation
Search	$t$ Stat	$p$ Value	Result
Snippet query-based is better than full sentence hybrid summary	1.76175	0.0802	H5: marginal confirmation

Table VIII. Generic Versus Other Summaries Overall

Overall	$t$ Stat	$p$ Value	Result
Generic is better than original summary	0.439075	0.6612	H6: rejected
Generic is better than snippet query-based summary	3.95	.00005	H6: supported
Generic is better than full-sentence hybrid summary	4.17	.00005	H6: supported

summaries contain slightly more page-level context than query-based summaries, which is of greater use in a browse scenario. Also, the full-sentence hybrid summaries utilized some discourse analysis to identify and select sentences from diverse topical areas, while the snippet, query-based summaries only considered similarity to the query terms as selection criteria. Likewise, our fifth hypothesis stated that the snippet, query-based summaries would outperform full-sentence, hybrid summaries in search tasks. The reasoning was that snippet summaries are more focused to the query and therefore are more useful given specific tasks, where users are less concerned about page-level context. The results from our statistical analysis (two-tailed  $t$  test) are listed in Table VII. Given a significance level of .10, both H4 and H5 were confirmed: the full-sentence hybrid summaries were more useful in browsing tasks ( $p$  value = .0678) and the snippet query-based summaries were more useful in search tasks ( $p$  value = .0802). The confirmation was only marginal, however, because the resulting  $p$  values did not reach the same .05 standard as did the values in the other statistical tests.

**8.3.4 Overall Summary Performance.** To test our sixth hypothesis, we statistically compared (using a two-tailed  $t$  test) the total number of relevant documents selected using the generic summary to the number selected by each of the other three types of summaries. Table VIII summarizes the statistical analysis for this comparison. While the generic summarizer outperformed both the hybrid and query-based summarizers ( $p$  values = .00005), the overall performance of the generic summary closely paralleled that of the original summary ( $p$  value = .661). Thus, we rejected our sixth hypothesis that the generic summary outperformed all other summaries overall.

**8.3.5 Time Spent on Summaries.** The time spent on summaries is also an important factor in evaluating their performance. We attempted to control for time by making the summaries the same length. In the experiment, however, subjects were not limited to the time they could spend on any one task, though they could not return to any task or lookup information not contained in the summary. We conducted some analysis on the time the users spent on each task given the type of summary. Table IX lists the average time spent by users given

Table IX. Differences by Summary Type on Time Spent per Task

Summary Type	Ave. Time per Task	$t$ Stat	$p$ Value
Time spent on generic summary different from the average (mean)	2.23	1.491	.137
Time spent on full-sentence hybrid different from the average (mean)	2.21	1.242	.215
Time spent on query-based snippet different from the average (mean)	2.02	1.12	.264
Time spent on original summary different from the average (mean)	1.58	1.61	.107
Time spent on browse tasks	2.09	—	—
Time spent on search tasks	2.13	—	—

each type of summary. Time did vary between tasks, but did not vary significantly from the mean time per task of 2 min and 11 s, as shown by the  $p$  values in Table IX. Original summaries showed the greatest deviation from the mean (13 s). Original summaries were typically shorter than the others, however, as we could not control their length. In addition, the time spent on browse tasks was also very similar to that spent on search tasks. The average time for browsing tasks was 2 min 9 s. The average time spent on searching task was 2 min 13 s. Because the average task time spent given different types of summaries did not vary significantly from the mean, time is not considered a contributing factor to better performance in the information-seeking experiment.

#### 8.4 Discussion of the Information-Seeking Summarization Experiment

In the following sections, we discuss the experimental results in terms of the importance of page-level context. We also discuss other findings of the information-seeking experiment.

**8.4.1 Summarization in Context.** The most useful text summary for Web searching depends on the user's task. Our primary finding is that, given a browsing task, users can better select relevant documents given a generic summary, while in a search scenario users make better relevance decisions using query-based and hybrid summaries. This finding is significant given the almost nonexistent use of generic summaries on the Internet today. This finding, however, is consistent with theories on information seeking. When users are more vague and uncertain about a search topic, as is more common in browse scenarios, their query formulation will be less precise. During this search stage, less focus should be placed on those query terms for summarization because a summary may or may not be what the users really want to see. The users are still in the process of refining their queries.

We have suggested the difference between the generic summaries and query-based summaries to be one of page-level context. Such difference arises because generic summaries commonly utilize discourse analysis and query-based summaries use term or entity-level analysis. Also, full-sentence summaries have more page context than snippet summaries. When a task is focused, as in a search task, there is less need for context and subsequently less robust discourse and language analysis. On the other hand, context becomes more

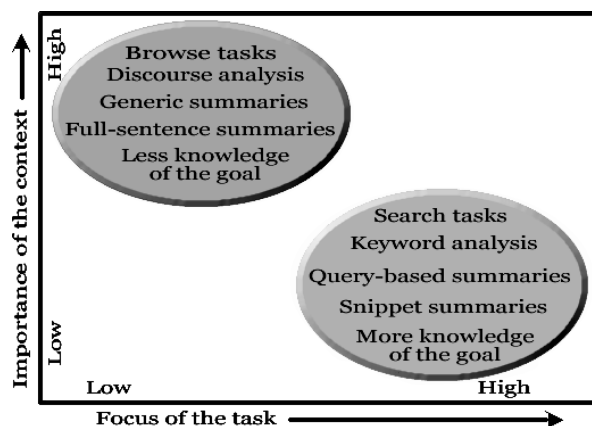


Fig. 4. Importance of the context given the focus of a task.

important when the task is less focused and more open ended. Figure 4 depicts the relationship between the importance of page-level context and the focus of the user’s task. The less focused a task is, the more important contextual information will be to help the user understand the document. Researchers have also characterized users’ feelings during the search process and found users browse more when uncertain about a topic [Kuhlthau 1991]. Given this connection in research, we have added “More knowledge of the goal” to Figure 4, even though we did not directly test users’ knowledge of the tasks given. This addition was made to show a connection between the knowledge levels characteristic of particular types of tasks. The addition also hints at future research that studies users’ familiarity with a topic, and how that familiarity affects the type of summary desired by users given a particular type of task.

Full-sentence summaries were more beneficial in browse tasks, while snippet summaries were more beneficial given search tasks. In snippet summaries, users are shown where the query terms are in the document more than in any other summary. Search tasks are well summarized by their query terms. In open-ended browse tasks, query terms insufficiently summarize what information is sought. Prominent page-level context may not be the information that is sought, but it does help users better judge the page’s overall relevance and refine their query terms for the next search.

The fact that query-based and hybrid summaries outperformed generic summaries in search tasks also provides some insight into the difference between search and browse tasks. The scope of information sought appears to be different. Searchers appeared less concerned with the relevance of an entire page and more concerned about whether they got the bit of information they were seeking. To users performing a search task, the boundaries between documents are less important. More information about the query terms is the desired information. Acquiring the needed information is the ultimate goal, and page-level context does not speed locating such information. Users that are browsing, however, are more interested in page-level context. Not being as familiar with the

topic, users who are browsing want to make sense of each page to help them reformulate their search and better understand their topic. For them, a generic summary along with its greater use of natural language processing provides the context needed to make relevance decisions.

*8.4.2 Implications for Information Retrieval.* The importance of page-level context for users when browsing has some implications for information system design. As mentioned in the introduction, users have been shown to distinguish between their searching and browsing behavior. With this knowledge, users could indicate the type of information-seeking task they are performing and the summaries displayed could adjust accordingly to best support their task. In addition, browsing tools could incorporate the distinction of query-related content in a document versus document-level context. Browsing tools may only include phrases from document-level context areas or tools may cluster pages based on similarity of the query-related content and similarity of the document-level context as separate dimensions in the clustering algorithm. Finally, ranking algorithms could rank higher documents that have the greatest overlap between query-related content and document-level context information. If a document's most prominent sentences (in a generic summarization sense) were also related to the query, then the document might have higher relevance.

*8.4.3 Native Versus Nonnative English Speakers.* While analyzing the results from the experiment, we also discovered some unanticipated differences between native and nonnative English speakers in performance using the snippet summaries. Nonnative English speakers did not find as many relevant documents using the query-based snippet summary as they did using the three other types of summaries compared to native English speakers ( $p$  value = .001). This finding was unanticipated, but appears to be consistent with our findings that page-level context is more helpful when users are less familiar with a topic or perhaps even with the language of the source material.  $p$  Values in a  $t$  test comparing native and nonnative English speakers were all above .10 for the use of generic, hybrid, and original summaries.

*8.4.4 Limitations of User Study.* In this research, we presented the same number of search tasks as browse tasks to our subjects. In doing so, we ignored the prevalence of one task over the other on the Internet. However, browsing is a common information-seeking process performed by users and, given such tasks, generic summaries more effectively guide users to relevant documents. In addition, a limitation of our study was the use of a single expert in evaluating the relevance of each document.

*8.4.5 Performance of Original Summaries.* The original summaries from the content providers on the whole performed well. In the search task, the original summary was statistically the best performer. In the browse task, the original summary was statistically better than the snippet, query-based summary and statistically the same as the full-sentence, hybrid summary. As mentioned, the original summaries contained a combination of human-generated,

lead sentence(s), and snippet summaries. The mix of summaries found in the original might account for the high and more balanced results in the experiment. A topic of future research will further explore combining both types of information, page-context and query-focused content, into one summary.

*8.4.6 Postquestionnaire Analysis.* After completing the experiment, users were asked to complete a short questionnaire. Users consistently chose a browse question as the most difficult and a search question as the easiest. This is interesting given that there was no statistical difference in performance or in time spent by users between browse and search questions in the experiment. Users were quite varied in their preference for different summaries. Overall, 34% of the users preferred the full-sentence, hybrid summary, 31% preferred the original summary, 22% preferred the generic summary, and 12% preferred the snippet, query-based summary. Participants who preferred the query-based summaries felt the computer was doing more work for them, which they liked. Those who preferred the original summaries liked how concise the summaries were. Finally, participants that preferred the generic summaries felt they were getting a better idea of what the document was really about.

## 9. CONCLUSIONS AND FUTURE DIRECTIONS

In this section, we summarize our findings as well as present future directions for this research.

### 9.1 Conclusions

We have developed a generic text summarizer, using a blend of discourse structural information as well as sentence-selection heuristics. The summarizer achieved performance equivalent to or better than two other published systems using the same corpus of human-generated summaries as the gold standard. The summarizer was then modified to create two additional types of summaries, one a full-sentence hybrid summarizer and the other a snippet, query-based summarizer. The hybrid summarizer used some discourse analysis and query-term information, while the query-based summarizer used only query-term similarity information.

Conferences on automatic text summarization including TIPSTER's SUMMAC in 1998 and the ongoing Document Understanding Conferences (DUCs) evaluate query-based summaries in information-seeking environments, while generic summaries are evaluated intrinsically or using nonsearching tasks. We have explored the use of four different summary types in search and browse information-seeking scenarios. Our findings indicate generic summaries are more useful than query-based summaries in open-ended tasks, despite their scarce use on the Internet. On the other hand, query-based summaries outperformed the generic summaries in more focused search tasks. We concluded page-level context helps users find more relevant documents when they are less focused in their task. Page-level context can be provided in summaries by not focusing on the query terms, using discourse-level information, and using full sentences as opposed to snippets in summary creation.

In highlighting the role of page-level context in information seeking, we utilized a very large sample size of 297 subjects. Such a large sample was important given the lack of research comparing generic and query-based summaries. This sample size is the largest we have seen exploring the use of summaries in information-seeking tasks.

## 9.2 Future Directions

We have compared the performance of summaries with high page-level context to those with high query focus. We are interested in testing summaries with both types of information together in an information-seeking experiment. Researchers in the field of information visualization have shown that “Focus + Context” interfaces [Plaisant et al. 1995; Greene et al. 2000] lead to faster user navigation [Pirolli et al. 2001; Börner and Chen 2002]. These interfaces deal with many documents or data records. We are interested in testing the “focus + context” paradigm in terms of single-document summaries that would provide some information focused on the query and other information relevant to the document-level context. The impact of “focus + context” in single-document summaries may help users make better relevance decisions and save time.

## APPENDIX. EXAMPLE SUMMARIES

### A.1 Generic Summary

ITworld.com—EDI? XML? Or Both?

*Summary:* Total **sentences:** 150; **top page terms:** xml, edi, technology, companies, connectivity. “One goal for exploring **XML** is to broaden our group of trading partners to include those who—for whatever reason—don’t use EDI,” says Ken Olsen, assistant vice president of marketing at Transentric. Thus, the overall process is a much bigger picture that this translation must fit into.

### A.2 Hybrid Summary

ITworld.com—EDI? XML? Or Both?

*Summary:* Total **query terms** in doc: 67; **which term(s)** found: xml. Many firms view the advent of **XML** as the golden opportunity to automate processes from beginning to end, with the **XML** format as the central touchstone. At the same time, such services typically use **XML** exclusively internally and communicate with **XML** users.

### A.3 Query-Based Summary

ITworld.com—EDI? XML? Or Both?

*Summary:* ...rtners. Now, along comes the document-tagging language **XML**, with the potential to reach new markets, simplify .....s. Should companies stay with EDI? Should they move to **XML**? Should they try to get EDI and **XML** to interoperate.....bove—if they use the right tools in the right ways. **XML** in Practice. Many enterprises embracing both EDI an...

## A.4 Original Summary

## ITworld.com—EDI? XML? Or Both?

*Summary:* Many organizations now using EDI are considering a move to XML to open up new markets. There are good tools to help, but you may want to hang onto at least some of your existing EDI links.

## ACKNOWLEDGMENTS

We would like to thank Kathy McKeown and Hongyan Jing for allowing us to use their TREC summarization data. We would like to thank Queen Booker for giving us access to her class and Ann Lally for her library science expertise. We also thank Jay F. Nunamaker, J. Leon Zhao, and the anonymous reviewers for their comments on earlier drafts of the article.

## REFERENCES

- AONE, C., HALVERSON, L., HAMPTON, T., AND RAMOS-SANTACRUZ, M. 1998. SRA: Description of the IE2 system used for MUC-7. In *Proceedings of the Message Understanding Conference-7*.
- AONE, C., OKUROWSKI, M. E., GORLINSKY, J., AND LARSEN, B. 1999. A trainable summarizer with knowledge acquired from robust NLP techniques. In *Advances in Automatic Text Summarization*, I. Mani and M. T. Maybury, Eds. MIT Press, Cambridge, MA, 71–80.
- BARZILAY, R. AND ELHADAD, M. 1999. Using lexical chains for text summarization. In *Advances in Automatic Text Summarization*, I. Mani and M. T. Maybury, Eds. MIT Press, Cambridge, MA.
- BLACK, E., JELINEK, F., LAFFERTS, J., MAGERMAN, D. M., MERCER, R., AND ROUKOS, S. 1992. Towards history-based grammars: Using richer models for probabilistic parsing. In *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*.
- BOGURAEV, B. AND KENNEDY, C. 1997. Saliency-based content characterization of text documents. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization at the ACL/EACL Conference* (Madrid, Spain).
- BÖRNER, K. AND CHEN, C. 2002. Visual interfaces to digital libraries. In *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries* (Portland, OR).
- BRANDOW, R., MITZE, K., AND RAU, L. F. 1995. Automatic condensation of electronic publications by sentence selection. *Inform. Process. Manage.* 31, 5, 675–685.
- BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertext Web search engine. In *Proceedings of the 7th International World Wide Web Conference* (Brisbane, Australia).
- CARBONELL, J. AND GOLDSTEIN, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR* (Melbourne, Australia).
- CARMEL, E., CRAWFORD, S., AND CHEN, H. 1992. Browsing in hypertext: A cognitive study. *IEEE Trans. Syst. Man Cybernet.* 22, 5, 865–884.
- CHEN, H., HOUSTON, A. L., SEWELL, R. R., AND SCHATZ, B. R. 1998. Internet browsing and searching: User evaluations of category map and concept space techniques. *J. Amer. Soc. Inform. Sci.* 49, 7, 582–603.
- CHEN, H., SCHUFELS, C., AND ORWIG, R. 1996. Internet categorization and search: A self-organizing approach. *J. Vis. Commun. Image Rep.* 7, 1, 88–102.
- CHOI, F. Y. Y. 2000. Advances in domain independent linear text segmentation. In *Proceedings of NAACL '00* (Seattle, WA).
- COVE, J. F. AND WALSH, B. C. 1988. Online text retrieval via browsing. *Inform. Process. Manage.* 24, 1, 31–37.
- EDMUNDSON, H. P. 1969. New methods in automatic extracting. In *Advances in Automatic Text Summarization*, I. Mani and M. T. Maybury, Eds. MIT Press, Cambridge, MA, 23–42.
- ETZIONI, O. 1996. The World Wide Web: Quagmire or goldmine? *Commun. ACM* 39, 11, 65–68.

- FIRMIN, T. AND CHRZANOWSKI, M. J. 1999. An evaluation of automatic text summarization systems. In *Advances in Automatic Text Summarization*, I. Mani and M. T. Maybury, Eds. MIT Press, Cambridge, MA, 325–336.
- GOLDSTEIN, J., KANTROWITZ, M., MITTAL, V., AND CARBONELL, J. 1999. Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*, 121–128.
- GREENE, S., MARCHIONINI, G., PLAISANT, C., AND SHNEIDERMAN, B. 2000. Previews and overviews in digital libraries: Designing surrogates to support visual information seeking. *J. Amer. Soc. Inform. Sci.* 51, 4, 380–393.
- HARTER, S. P. 1992. Psychological relevance and information science. *J. Amer. Soc. Inform. Sci.* 43, 9, 602–615.
- HEARST, M. A. 1997. Segmenting text into multi-paragraph subtopic passages. *Computat. Linguist.* 23, 1, 33–64.
- HEARST, M. 1999a. User interfaces and visualization. In *Modern Information Retrieval*, R. Baeza-Yates and B. Ribeiro-Neto, Eds. ACM Press, New York, NY, 257–339.
- HEARST, M. A. 1999b. User interfaces and visualization. In *Modern Information Retrieval*. Harlow, UK, Addison Wesley.
- HEARST, M. A., ELLIOT, A., ENGLISH, J., SINHA, R., SWEARINGEN, K., AND YEE, K.-P. 2002. Finding the flow in Web site search. *Commun. ACM* 45, 9, 42–49.
- HERSH, W., PENTECOST, J., AND HICKEM, D. 1996. A task-oriented approach to information retrieval evaluation. *J. Amer. Soc. Inform. Sci.* 47, 1, 50–56.
- JING, H., BARZILAY, R., MCKEOWN, K., AND ELHADAD, M. 1998. Summarization evaluation methods: Experiments and analysis. In *Proceedings of the AAAI Spring Symposium on Intelligent Text Summarization*, 51–59.
- JURAFSKY, D. AND MARTIN, J. H. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, NJ.
- KAN, M.-Y., KLAVANS, J. L., AND MCKEOWN, K. R. 1998. Linear segmentation and segment significance. In *Proceedings of the 6th International Workshop of Very Large Corpora (WVLC-6: Montreal, P. Q., Canada)*, 197–205.
- KATZ, S. M. 1996. Distribution of content words and phrases in text and language modelling. *Nat. Lang. Eng.* 2, 1, 15–59.
- KLEINBERG, J. 1999. Authoritative sources in a hyperlinked environment. *J. Assoc. Comput. Mach.* 46, 5, 604–632.
- KUHLTHAU, C. C. 1991. Inside the search process: Information seeking from the user's perspective. *J. Amer. Soc. Inform. Sci.* 42, 5, 361–371.
- KUPIEC, J., PEDERSEN, J., AND CHEN, F. 1995. A trainable document summarizer. In *Proceedings of the 18th ACM-SIGIR Conference*, 68–79.
- LAM-ADESINA, A. M. AND JONES, G. J. F. 2001. Applying summarization techniques for term selection in relevance feedback. In *Proceedings of SIGIR (New Orleans, LA)*.
- LANDAUER, T., EGAN, D., REMDE, J., LESK, M., LOCHBAUM, C., AND KETCHUM, D. 1993. Enhancing the usability of text through computer delivery and formative evaluation: The SuperBook project. In *Hypertext: A Psychological Perspective*, C. McKnight, A. Dillon, and J. Richardson, Eds., Ellis Horwood, Chisherten, U.K., 71–136.
- LIDDY, E. D. 1998. Enhanced text retrieval using natural language processing. *Amer. Soc. Inform. Sci. Bull.* 24, 4 (April/May), 14–16.
- LUHN, H. P. 1958. The automatic creation of literature abstracts. In *Advances in Automatic Text Summarization*, I. Mani and M. T. Maybury, Eds., MIT Press, Cambridge, MA, 15–22.
- MANI, D., HOUSE, D., KLEIN, G., HIRSCHMAN, L., OBRST, L., FIRMIN, T., CHRZANOWSKI, M., AND SUNDHEIN, B. 1998. The TIPSTER SUMMAC text summarization evaluation: Final report, DARPA, tech. rep. Defense Advanced Research Projects agency, Arlington, VA.
- MANI, I. AND MAYBURY, M. T. EDS. 1999. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA.
- MARCHIONINI, G. AND SHNEIDERMAN, B. 1988. Finding facts vs. browsing knowledge in hypertext systems. *Computer* 21, 1, 70–79.

- MARCU, D. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA.
- MARSHALL, B., McDONALD, B., CHEN, H., AND CHUNG, W. 2004. EBizPort: Collecting and analyzing business intelligence information. *J. Amer. Soc. Inform. Sci. Tech.* 55, 10, 873–891.
- McDONALD, D. AND CHEN, H. 2002. Using sentence-selection heuristics to rank text segments in TXTRACTOR. In *Proceedings of the Second ACM/IEEE-CS JCDL* (Portland, OR).
- McDONALD, D. M., CHEN, H., SU, H., AND MARSHALL, B. B. 2004. Extracting gene pathway relations using a hybrid grammar: The Arizona Relation Parser. *Bioinformatics* 20, 18, 3370–3378.
- MINEL, J.-L., NUGIER, S., AND PIAT, G. 1997. How to appreciate the quality of automatic text summarization. In *Proceedings of the ACL/EACL '97 Workshop on Intelligent Scalable Text Summarization*.
- MORRIS, A., KASPER, G. M., AND ADAMS, D. A. 1992. The effects and limitations of automatic text condensing on reading comprehension performance. *Inform. Syst. Res.* 3, 1.
- NOMOTO, T. AND MATSUMOTO, Y. 2001. A new approach to unsupervised text summarization. In *Proceedings of the SIGIR* (New Orleans, LA).
- OVER, P. AND YEN, J. 2004. An introduction to DUC 2004: Intrinsic evaluation of generic news text summarization systems. In *Proceedings of the Document Understanding Workshop, HLT/NAACL* (Boston, MA).
- PIROLI, P., CARD, S. K., AND VAN DER WEGE, M. M. 2001. Visual information foraging in a focus + context visualization. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 508–513.
- PLAISANT, D., CARR, D., AND SHNEIDERMAN, B. 1995. Image-browser taxonomy and guidelines for designers. *IEEE Softw.* 12, 2, 21–32.
- RADEV, D. R., JING, H., AND BUDZIKOWSKA, M. 2000. Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the ACL/NAAL Workshop on Summarization* (Seattle, WA).
- SANDERSON, M. 1998. Accurate user directed summarization from existing tools. In *Proceedings of the Conference on Information and Knowledge Management* (Bethesda, MD).
- SCHAMBER, L., EISENBERG, M. B., AND NILAN, M. S. 1990. A re-examination of relevance: Towards a dynamic, situational definition. *Inform. Process. Manage.* 26, 6, 755–776.
- STRZALKOWSKI, T., WANG, J., AND WISE, B. 1998. A robust practical text summarization. In *Proceedings of the AAAI-98 Spring Symposium on Intelligent Text Summarization*, 26–33.
- TEUFEL, S. AND MOENS, M. 1999. Sentence extraction as a classification task. In *Proceedings of the Workshop on Intelligent Scalable Summarization, ACL/EACL Conference* (Madrid, Spain).

Received November 2003; revised June 2004, September 2004; accepted November 2005