**RC22719 (W0302-023) February 5, 2003**
**Computer Science**

# IBM Research Report

## Comparing Search Results Summaries for Technical Support Documents: A User Study

**Catherine G. Wolf, Sherman R. Alpert, Lev Kozakov, John G. Vergo, Yurdaer N. Doganata, Clare-Marie Karat**
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

**Research Division**
**Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich**

# Comparing Search Results Summaries for Technical Support Documents: A User Study

**Catherine Wolf**
cwolf@us.ibm.com

**Sherman Alpert**
salpert@us.ibm.com

**Lev Kozakov**
kozakov@us.ibm.com

**John Vergo**
jvergo@us.ibm.com

**Yurdaer Doganata**
yurdaer@us.ibm.com

**Clare-Marie Karat**
ckarat@us.ibm.com

**IBM TJ.Watson Research Center**

## ABSTRACT
Pressure to reduce costs has driven many businesses to move technical support from human help desks to the Web. Customers expect to find information easily and quickly when they visit a support site. If they don't, they may turn to a competitor. This paper reports on an empirical evaluation of different types of summaries for the presentation of search results on the technical support website of a large computer manufacturer. The summary types tested were: the summaries shown on the current live site, our own rule-based abstracts, a search terms in context summary, and no summary at all (that is, titles only). Although there were no significant differences between summary types on most measures, the overall pattern of results suggested that the rule- based abstracts were better than the titles only. Notably, 30% of the people in the abstract group were often able to select the right document based solely on information contained in the summary. The implications for summarizing technical support documents for search are discussed.

## Categories and Subject Descriptors
H.3.1 [INFORMATION STORAGE AND RETRIEVAL]: Content Analysis and Indexing---*Abstracting methods*; H.3.3 [INFORMATION STORAGE AND RETRIEVAL]: Information Search and Retrieval---*Search process, Selection process*

## General Terms
Human Factors

## Keywords
user studies, Web search, document summaries

## 1. INTRODUCTION
Ten years ago, when customers needed technical support, they would pick up the phone and call an expert at a help desk. Today the cost of providing support to customers, partners and employees is placing an increasing burden on corporate profitability. The pressure to reduce costs combined with the rapid growth of the Web has caused companies to move support from human experts to the Web. Many companies charge for this formerly free human service, motivating customers to seek assistance online. As Ehrlich has noted [4], the skills of a help desk organization may be hard to replicate with online self service, but the increase support cost will force many companies to look at cost effective solutions over the Web.

Coupled with the need to reduce costs, is the often conflicting goal of increasing customer satisfaction. The quality of support and service can be the differentiator that sets one business apart from its competitors. When customers go to a business' website seeking information or the solution to a problem, they expect to find it quickly and with a minimum of effort. If they don't, they may turn to another supplier. Thus, it is important for a business to develop and maintain customer loyalty by ensuring that customers achieve their goals when they search for information on its technical support website.

There are many aspects of providing high quality technical support on the Web. One factor that affects whether customers are able to find the information they need is the quality of the summaries of documents retrieved as the result of a search. A large business may have tens of thousands of support documents, written by different authors and not necessarily originally intended for customers. Creating effective summaries manually is an expensive proposal, especially for organizations that have huge repositories containing document without abstracts. Therefore, summaries must be created programmatically rather than by hand.

There are a number of approaches to generating summaries of documents. Information on automatic summarization techniques and products may be found in [8], additional bibliography and other resources may be found in [9] and [10]. Today there exist several intelligent summarization systems [10] that combine various methods of natural language processing with well known librarian techniques [1, 6].

The search engines on the Web have several different approaches to summary creation. The major Web search gateways, like Google.com, AllTheWeb.com or AltaVista.com find snippets of text that contain the user's search terms in a document and display these snippets as the summary with highlighted terms. This method, sometimes called "terms highlighted in context" (THIC), is also used by some corporate technical support search sites, like Hewlett-Packard (http://search.hp.com/) and Apple.com (http://kbase.info.apple.com/). Some web search engines, like Go.com (http://search.go.com/) use metadata in the HTML


*Conference '00*, Month 1-2, 2000, City, State.

encoding of a document—for example, many document authors include a "description" metatag, and that tag's value is used by some search engines as the document's summary. Other engines use the first *n* characters or words of the body of a page as the page's summary, or use a hybrid approach utilizing several of these techniques. Some well-organized corporate technical support sites, like Microsoft (http://search.support.microsoft.com/), use nice human created summaries that are included in the documents.

This paper reports on an empirical evaluation of different types of summaries for the presentation of search results on the technical support website of a large computer manufacturer. The purpose of this evaluation is to determine what type of summarization techniques are more likely to help customers find information accurately and quickly and, hence, increase customer goal attainment and, thus, customer satisfaction. The study of the effectiveness of different summary types was part of an effort to improve the user experience on the website. In the rest of the paper, we describe the website and documents, the summary conditions tested including a high level description of the algorithms used to generate summaries, the design and methods of the study and the results. We conclude with a discussion of the findings and the implications for the creation of technical support document summaries on the Web.

## 2. WEBSITE AND DOCUMENTS

The website that is the subject of this evaluation provided technical support for a large range of hardware and software products including everything from laptop computers to servers to point of sale systems, as well as application software and developer tools. There were about 500,000 documents on the website.

The documents on the support site are not homogeneous with regard to content or format; rather, there are many different document types. Documents of different types and even documents of one particular type may be created by different organizations within the enterprise, sometimes with differing tools. The result is a lack of consistency across and within document types with regard to document content, subcomponents, and format. There were about 16 major document types with a number of variants within many types. The types included downloads, FAQs, white papers, product information, problem reports, and hints, among others. The technical complexity ranged from highly technical to basic.

The users of the site included both information technology specialists and home and home office users.

## 3. SUMMARY TYPES

There were 4 summary types tested in the study. These were: the summaries shown on the current live site, our own rule-based abstracts, a search terms in context summary, and no summary at all (that is, titles only). In the rest of this paper, these will be identified as *Live site*, A*bstract*, *THIC (term hits in context)*, and T*itles only*.

### 3.1 Live Site

Live site summaries were obtained by simply performing actual searches on the live site and caching the results. Interestingly, the summaries from the live site exhibit a broad range of quality and type. Some were all or part of a section of the document. Others consisted of the first 255 characters of the body of the document, and a few consisted of the title of the document. When the summary was the first 255 characters of the body, it could contain headings such as "Applicable countries." For a small number of documents, the first 255 characters were a standard message that gave no clue to the document's contents. Because of the diversity of summary quality and content, the website owners were looking for a way to improve the summaries.

### 3.2 Abstract

The process for creating the rule-based abstracts was a bit more involved. We intended our abstract to be composed of specific pieces of the actual document—that is, we wished to programmatically craft an abstract using text from the document itself. We did have the option of employing an intelligent programmatic summarization tool, and found that it performed better when analyzing a particular section of a document than when attempting to summarize the entire document contents. The authors analyzed each document type and its multiple subcomponent sections and multiple variants to determine which pieces could be used (as is or summarized) for a cogent and "good" summary. In a separate preliminary study we also asked document authors to compose summaries for a sample of documents using whole sentences from the documents. This analysis led to a set of imperative and conditional rules for the composition of abstracts for each document type. For example, even though "Frequently Asked Questions" documents share the same document type, there exist many different formats for these documents. The following describes the process for composing the abstract for FAQ documents (the notation "Summarizer(<section-name>)" means get all the text of the section labeled "<section-name>" and invoke the programmatic summarizer to obtain a summary of that text):

- Output "[FAQ] "
- If there exists a section labeled "Number:"
  - if there is no "Problem" section in this document, then abort this process and summarize this document the same as a "White Paper" document (see "White Paper" rules)
- else if there is a "Problem" section, output Summarizer(Problem)
- else if there is an "Abstract" section, output Summarizer(Abstract)
- else if there is a section title that begins with "Question" (that is, "Question" or "Question**:**") then
  - if the contents of that "Question/Question**:**" section are not the same as the title, output Summarizer(that section)
  - else if there is a "Question**:**" (with colon) subsection in the "Answer" section, output Summarizer(Question**:**)
- Output the "Product Information" section (with predetermined product-info format)

The programmatic summarizer, used in the study, implements a well known sentence extraction model [5, 7] analyzing lexical cohesion factors in the source document text [2]. Sentence extraction is driven by the notion of salience - the resulting

summary is constructed by identifying and extracting the most salient sentences in the source document. The salience score of the sentence is defined partly from the salience of the vocabulary items in it, and partly from its position in the document structure and the salience of surrounding sentences.

## 3.3 THIC

For THIC summaries we also coded a programmatic solution. In a real world context, THIC summaries would be driven by the actual terms entered by the user. In our study the search terms were not selected by the participants, but chosen in advance for each search task by the authors (see Method section for details of how the search terms were chosen and the rationale for choosing them in advance).

Each document summary in the results list contains one or more text snippets, each illustrating an instance of the use of one or more search terms in the document. In the simplest case, each snippet contains a contiguous chunk of text from the document in which a particular search term is shown along with surrounding text, that is, a fragment of text before the search term, the search term, and a fragment of text after the term. For example, the search terms "java" and "text" might result in snippets "A primary design goal of **Java** is to allow developers to write software that can …" and "... documentation regarding writing a **text** editor application in ..." Our simplified algorithm begins by finding the first occurrence of each search term in the document, and for each extracting a text snippet (of approximate length of 155 characters) showing the term in context. Then overlapping snippets are merged thereby illustrating snippets wherein more than one search term occurs—that is, if the first snippet containing "Java" overlaps with the first "text" snippet, the two snippets are merged into one, with additional processes to minimize the length of the resultant snippet. Merging is performed recursively on all resultant snippets (which becomes more important when there are more than two search terms). Snippets are always extended at the "edges" (the head and tail) so as not to truncate words. Lastly, if the front of the snippet does not occur at the beginning of a sentence, ellipses are prepended to the head of the snippet; similarly, if the tail of a snippet is not the end of a sentence, ellipses are appended to the tail.

## 3.4 Titles Only

The Titles only condition used the titles of the documents from the live site with no additional information.

## 3.5 Other Features

An additional objective was to learn if three other features were helpful to users in deciding which documents were relevant to their needs. These features were bolding of search terms, inclusion of document type information, and inclusion of formatted product information in technical support document summaries. The search terms were bolded in the THIC summaries, and also in Abstract summaries when present. Product information was taken from the documents and included such information as product line, operating system, hardware platform, and version. The precise information included depended on the product. The Abstract and THIC conditions had product information. The Abstract condition had document type information as well.

Examples of the three document summary types are shown in the Appendix.

## 3.6 Hypothesized Advantages of Different Summary Types

The Live site summaries are a mixed bag. We expected that those that were all or part of a human authored section to fare well if the section chosen matched the information needs of the task. Target documents whose summaries consisted of the first 255 characters were unpredictable; the results would depend on the contents of the characters. Because Abstract summaries were composed of sentences from the document, we thought they might be easier to read than the THIC summaries which were composed of disconnected snippets of text. Also, if the rules for selecting sections to be summarized and the summarizer's notions of salience were correct, the Abstract summaries might do a good job of distilling the core content of documents. On the other hand, THIC summaries had the benefit of containing the search terms, which presumably got to the heart of the tasks. The quality of THIC summaries would depend on the search terms used. To avoid biasing the results of the study for or against THIC, we took the search terms from a list of popular terms for the site and chose target documents from among the 10 most frequently accessed documents for the terms in most cases. We could anticipate no advantage for the Titles only condition.

# 4. METHOD

## 4.1 Study Design

Different groups of people served as participants in each of the 4 summary conditions. We used 8 document types as targets (the document that was the right answer to the task). Each person got the same 24 tasks, 3 for each document type. The tasks consisted of brief scenarios that required the participants to find the solution to a problem or information on a topic. There were 10 results per search task and the target appeared about equally often in each position. We divided the 24 tasks into 3 sets so that each document type appeared as the target once in each set. Task order was counterbalanced across participants by varying the set according to a Latin square.

## 4.2 Participants

The participants were recruited by a market research agency. All were information technology (IT) professionals who were users of the website. All had used the site once a month or more in the past 6 months and in the past year had gone to website to seek information about laptop or desktop computers, basic computer software such as email, and complex hardware or software such as servers or management information systems. We chose to exclude home users because IT professionals were the majority of the site's users and accounted for the bulk of the business' revenue. There were 10 subjects in the Abstract group, 9 in Live site, 9 in THIC, and 10 in Titles only.

## 4.3 Materials

In order to have documents and search terms that were representative of those used by real users, we took most of the terms and documents from a list of the 250 most frequent search terms and the top 10 documents clicked on for each term for a recent month. Nine tasks had a 1-word search term, 8 had a 2-word term, 5 had a 3-word term, and 2 had a 4-word term.

To create the scenarios, we started with the search terms and documents and developed plausible tasks that could be answered by the documents. For example, one scenario was "You are using

the Sun Java Development Kit but you need to move the one provided by Microsoft. You're looking for an easy way to convert your applications to the Microsoft JDK." The search terms were "java conversion." One document, the target, contained the answer to the task. The other non-target documents were obtained by searching the live site with the search terms.

## 4.4 Procedure

Participants interacted with a software tool constructed for the study. This tool presented each search scenario to participants, performed the (simulated) search for documents matching a set of search terms, and displayed the search results using one of the four formats as set by the experimenter. The associated search terms for the task were displayed by the program in the search-term entry field. We provided a fixed set of search terms for each task, rather than having each participant enter terms of her choosing, for a number of reasons. First, we were not testing participants' query formulation skills, but rather how the content and format of the summaries shown in the results list affected the ability to find relevant and useful documents. Second, all of the search results pages were created a priori and cached in order to ensure that the system response time was equal for all 4 summary conditions. If the results list had been obtained on the fly from the live site, the extra processing needed to create the other 3 summary types would have resulted in longer system response times for these conditions.

After participants viewed the task scenario on the screen, they pressed the "Search" button. In response the study tool displayed the results list. Each entry in the results list consisted of a number for its position in the list, the title, which was a hyperlink to the document, and the summary (except in the Titles only condition). All hyperlinks were live; when a document link was clicked, the actual document was displayed in a standard browser window.

We told participants to find the document that contained the information needed to resolve the scenario. We asked them to treat this as a "real world" problem and to give it the same diligence as they would a similar problem in their actual work environment. The participants could click on links to the documents to see the documents. When they thought they had found the right document, they clicked the "Done" button and entered the number of the chosen document. We logged and time-stamped the start of search, the choice of right document, and all documents clicked on for each task.

# 5. RESULTS

Due to occasional problems with the live site, the study tool, and participants' behavior, not all participants had 24 valid tasks. Accordingly, in this section we report mean number correct per task rather than mean absolute number correct. Although there were statistically significant differences due to target document type, we report data averaged across target type. Given the variety of document formats and content within a document type, we could make no meaningful conclusions about document type with only 3 examples per type. There was no interaction between target type and summary condition. The mean number correct for the 4 summary conditions and 8 document types is shown in Figure 1.

## 5.1 Mean Number Correct and Mean Time

The mean number correct and the mean time to complete correct tasks is given in table 1 The time to complete tasks was measured from the time that participants clicked the search button to the time they clicked "Done. ".We used 2 scoring criteria; one we called strict which required the document that we had chosen (or in 1 case there was a second document that clearly was appropriate). For the other criterion, the liberal criterion, we counted as correct any document that 9 or more people chose for the task. We verified that these were reasonable choices.

A one way analysis of variance was done for the 4 measures in table 1. With both criteria the same pattern of results emerges. The differences among means for number correct and mean time are not statistically significant. Note that there is a speed-accuracy tradeoff for the 3 types of summaries (Abstract, Live site, THIC) The means for error rate and speed are oppositely ordered.

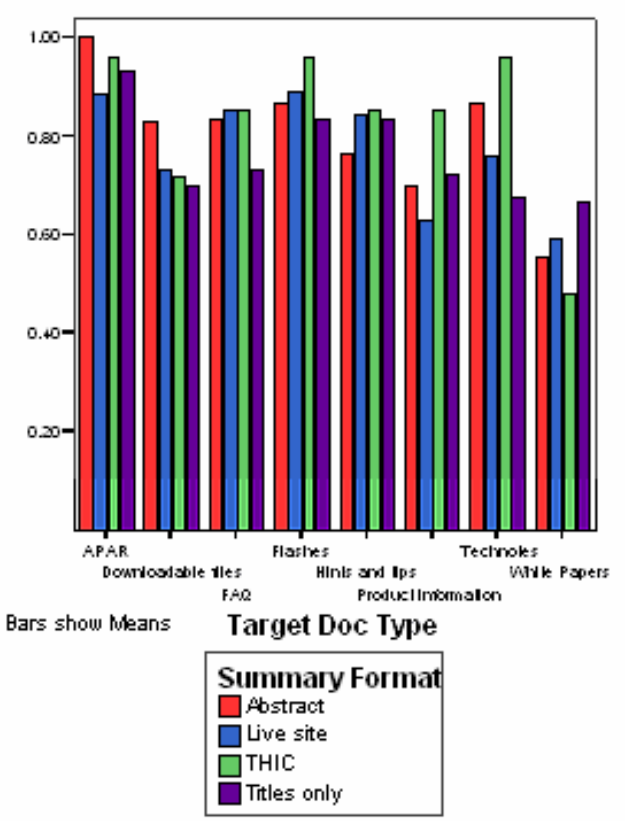


**Figure 1. Mean number correct for the 4 summary conditions and 8 target document types.**

**Table 1. Mean number correct per task and mean task completion time for correct tasks, strict and liberal criteria**

| Summary | Mean number correct, strict | Time for correct, strict (sec) | Mean number correct, liberal | Time for correct, liberal (sec) |
|---|---|---|---|---|
| Abstract | .80 | 92 | .87 | 96 |
| Live site | .77 | 84 | .84 | 86 |
| THIC | .83 | 114 | .88 | 115 |

| | | | | |
|---|---|---|---|---|
| Titles | .76 | 87 | .82 | 88 |

Because of the variability and skewness of response time data (typically the distribution has a long tail to the high side), we applied a commonly used transform, a log transform, to the mean time for the strict criterion. The results of analysis of variance on log mean time yielded no significant differences.

## 5.2 Mean Number of Clicked Documents

Another measure of the effectiveness of a summary type is the number of documents that people clicked for a task. The more informative a summary type is, the fewer documents people will need to read to find the right document. The mean number of clicked documents for correct tasks and all tasks is shown in table 2.

**Table 2. Mean number of clicked documents for correct tasks, strict criterion, and all tasks**

| Summary | Mean number clicked for correct | Mean number clicked for all |
|---|---|---|
| Abstract | 1.44 | 1.77 |
| Live site | 1.72 | 2.05 |
| THIC | 1.78 | 1.93 |
| Titles | 1.97 | 2.31 |

Although the differences among these means are not statistically significant, the Abstract group had the lowest number of clicked documents and the Titles only group had the highest. Three people in the Abstract group had mean number of clicks for correct tasks less than 1. One person in the Titles only group had a mean less than 1;nobody in the other 2 groups had a mean less than 1. The mean number of documents clicked for correct tasks and the mean number correct, as well as mean time for the 3 participants in the Abstract group, are given in table 3.

**Table 3. Mean number of clicked documents for correct tasks, mean number correct, and mean time for correct tasks, strict criterion, for the 3 people in Abstract group with mean number of clicked documents less than 1.**

| Mean number of clicked documents for correct | Mean number correct | Mean time for correct tasks (sec) |
|---|---|---|
| .39 | .75 | 66 |
| .70 | .83 | 61 |
| .10 | .83 | 104 |

These people were often able to select the right document based only on the information provided in the summary and did so with the same accuracy as the average. Two had considerably lower times than average and 1 was statistically indistinguishable from the mean for the Abstract group.

In contrast, the participant in the Titles only with mean number of clicked documents less than 1 had the highest error rate of the Titles only group and the lowest time for correct tasks. Unlike the 3 people in the Abstract group with mean number of clicked documents less than 1, this person sacrificed accuracy in clicking on few documents.

## 5.3 Predictors of Number Correct

In order to gain some insight into the factors related to the number correct, we did a multiple linear regression using time for correct tasks, time for all tasks, and number of clicked documents for correct tasks. Table 4 shows Pearson multiple R, which represents how well all factors combined predict number right, and the partial correlations of each factor with number right, which are the correlations of each factor with number right when the effects of all other factors have been removed.

**Table 4. Multiple regression predicting number correct and partial correlations for each summary type and overall**

| Summary | R | Time for correct (partial) | Time for all (partial) | Number clicked docs (partial) |
|---|---|---|---|---|
| Abstract | .86 | .71 | -.53 | -.11 |
| Live site | .30 | .29 | -.30 | .15 |
| THIC | .65 | .61 | -.57 | -.14 |
| Titles | .68 | .47 | -.38 | .11 |
| Overall | .63 | .48 | -.36 | -.05 |

The highest R is for the Abstract group, which means that the time for correct tasks, time for all, and number of clicked documents predicts the number right better for this group than the others. The partial correlation of number right with time for correct tasks for the Abstract group is the largest of the partial correlations. This means that the relationship between the 3 predictors and number right is mainly due to time for correct tasks. People in the Abstract group who took longer on correct tasks tended to do better. The same general pattern of results holds for THIC and Titles only groups. To aid in understanding the multiple regression results, the correlations between measures for the Abstract group is shown in table 5.

**Table 5. Correlations between measures for Abstract group**

| | Number right | Time for correct | Time for all | Number clicked docs |
|---|---|---|---|---|
| **Number right** | 1 | .80 | .69 | .24 |
| **Time for correct** | .80 | 1 | .97 | .42 |
| **Time for all** | .69 | .97 | 1 | .44 |
| **Number clicked docs** | .24 | .42 | .44 | 1 |

As can be seen from table 5, number of clicked documents is moderately correlated with time for correct tasks, but has a low correlation with number right.

## 5.4 Categorization of Summary Types Within Live Site

As we pointed out, the summaries on the Live site were not homogeneous but rather, consisted of 3 types: the first 255 characters of the body of a document, all or part of a section of a document or the title. We further divided the first category into summaries that contained no information about the document contents, and those that contained relevant information. In order to see if the subtype of summary for a target document on the Live

site affected accuracy, we grouped the targets into the 4 categories. There were 3 of the first 255 characters type containing no information (nb in Figure 2), 6 of the first 255 characters type with relevant information (ng), 10 of the section type (s), and 5 of title type (t). Figure 2 shows the mean number correct for these 4 categories for the Live site and the other summary conditions for comparison.

Looking first at the Live site, it's surprising that accuracy is best for the title type. But looking at the other summary conditions, we see that the title type had the highest accuracy in the other summary conditions as well. Furthermore, the pattern of accuracy is similar for all summary conditions. Since this classification is based on the subtypes within the Live site condition, the pattern of results must not be due to the summaries. For example, the THIC condition had summaries composed of snippets of text with search terms. Thus, these summaries always had information about the document, yet accuracy was lowest for THIC for the same tasks as the Live site condition- those corresponding to the first 255 characters with no document information in the Live site condition. Some possible explanations for these findings will be presented in the Discussion section.

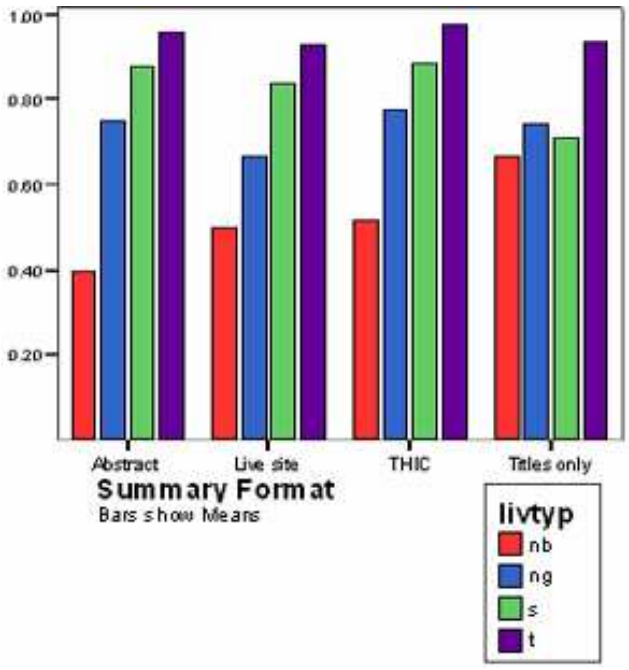


**Figure 2. Number correct, strict criterion, based on categorization of subtypes within the Live site for the 4 summary conditions. nb=first 255 char with no information about document, ng=first 255 char with document information, s=all or part of a section, t=title.**

## 5.5 Helpfulness of Information and Confidence in Choice

After every 4 tasks people were asked the following questions:

- For the task just completed, how helpful was the information in the search results list?
- For the task just completed, how confident are you that you found the right document?

For the first question a 7 point rating scale was used going from "very helpful" to "not at all helpful", with 1 being the most favourable rating. For the other question, the rating was made on a 7 point scale going from "very confident" to "very unconfident." We asked the same questions at the end of the session for the tasks overall. Table 6 gives the mean ratings for these questions for the 4 summary types.

For the task just completed, the rating for confidence for the Abstract group is marginally significantly better than the rating given by the Titles only group. Although the other differences are not statistically significant, the trend is clear. The Titles only has the worst rating in 3 out of 4 cases. Abstract always has the best rating.

**Table 6. Mean ratings for questions about helpfulness of information and confidence in choice. Means with * are significantly different at the p< .06 level.**

| Sum-mary | How helpful informa-tion last task | How confident last task | How helpful informa-tion overall | How confident overall |
|---|---|---|---|---|
| Abstract | 1.98 | 1.83* | 2.10 | 2.20 |
| Live site | 2.41 | 2.44 | 2.44 | 2.44 |
| THIC | 2.41 | 2.15 | 2.63 | 2.33 |
| Titles | 2.77 | 2.87* | 2.90 | 2.40 |

For the task just completed, the rating for confidence for the Abstract group is marginally significantly better than the rating given by the Titles only group. Although the other differences are not statistically significant, the trend is clear. The Titles only has the worst rating in 3 out of 4 cases. Abstract always has the best rating.

## 5.6 Bold Search Terms, Product Information, and Document Type

After they had finished all tasks, we asked participants about the helpfulness of product information, having the search terms in bold, and the document type. The means are shown for the groups to which these questions applied in table 7. The same 7 point scale was used, with 1 the most positive rating. T-tests were used to evaluate the statistical significance of the differences between means for bold search terms and product information. There were no significant differences between means. All features received moderately favorable ratings.

**Table 7. Mean ratings for product information, bold search terms and document type.**

| Summary | Product | Bold | Doc type |
|---|---|---|---|
| Abstract | 2.90 | 3.50 | 2.11 |
| Live site | N/A | N/A | N/A |
| THIC | 2.89 | 2.33 | N/A |

Some more insight into the value of these features comes from comments participants made. A number of people said that operating system information was the most important product

information. Several also said it was useful to separate hardware from software problems from drivers. A number said that drivers were the most important document type.

### 5.7 Other Comments

People made a variety of comments when asked how search could be improved and what they liked about other sites. The comments have to do with both document content and search features. Many said that documents, summaries, and titles should be in clearer language, not "technogorp." Titles should be more descriptive. It should be easier to see at a glance what a document contains.

Many users said they liked Google and wanted features that it has. People mentioned specifically that they wanted to search within results and refine the search. Some said that search should be more "intelligent. " For example, show results that may not match the search terms but have the same meaning. Another said that if he is in a particular product or category and then clicks "drivers" he should get drivers for that product, not a general driver search. Some stated that other sites track which products they own and use that information in search.

## 6. DISCUSSION AND IMPLICATIONS FOR TECHNICAL SUPPORT ON THE WEB

In this study, the Titles only and Live site conditions resulted in shorter times and lower accuracy, but none of the differences were not statistically significant. There appears to be a speed-accuracy tradeoff in that the Titles only and Live site groups had the worst accuracy and THIC the best, but for time to complete correct tasks the order was reversed. In general, the summary for the Live site was shorter than for the Abstract or THIC. Titles only had no summary and so had the briefest information in the results list. This probably partly explains the finding that accuracy was lowest for Titles only and the Live site but speed was fastest. The speed-accuracy tradeoff was evident within the groups except for the Live site in that people who had longer times for correct tasks tended to get more right (the Live site group had a small positive correlation). It should be remembered, however, that the differences between means for accuracy and time were not statistically significant.

We also looked at the average number of documents clicked for correct tasks. These results favored the Abstract, in that 3 subjects in this group had an average of less than 1. Their accuracy was about the same as the mean. In other words, the 3 people in the Abstract group were often able to choose the right document using only the information in the summary without sacrificing accuracy. The multiple regression analysis confirmed that the number of clicked documents was not correlated with accuracy for any group. These 3 also had task completion times that were faster or the same as the mean for the group. The number of clicked documents had a moderate positive correlation with time for correct tasks, suggesting that people who clicked more documents had longer task times. The Titles only group had the highest number of clicked documents, but the differences between means were not statistically significant. In real life people would probably spend more time than in the study looking at each document clicked, so the time savings for looking at fewer documents might be greater.

People in the Abstract group were also marginally more confident that they had found the right document than those in the Titles only group. The Abstract group also had the best ratings on all 4 judgments of helpfulness and confidence and the Titles only had the worst on 3 out of 4, but again the differences were not significant. The better ratings of the Abstract group might translate into greater user satisfaction.

Given the overall pattern of results, we can probably conclude that Titles only is inferior to Abstract. The success rate in this study was higher than for the actual website due to the need to create scenarios that could be described in a few sentences, whose goals were clear and had a solution in the set of documents retrieved. With lower success rates, it's possible that the differences found in this study might be magnified.

It is interesting that the participants in Titles only group did as well as they did. We observed in the other groups that many people seemed to scan the titles before deciding which summaries to read. That is probably an efficient strategy because it allows the user to eliminate many candidates without reading the summaries. The implication for technical support documents and documents on the Web, in general, is that the title should be highly descriptive of the content.

Given that the THIC approach seems to be the method of choice for general web search engines, it is surprising that it fared no better than the other methods. Also THIC summaries presumably contained relevant portions of the document since the text was centered on the search terms for a task. Creating summaries of documents for a technical support website, however, is different from the task of creating summaries for any and all documents on the Web. Although the structure of these technical documents was varied, it was known and the Abstract method could use rules that took advantage of this structure. In contrast, general web search engines must handle unpredictable structure, and a single document may contain many topics of no interest to the person searching. Perhaps for technical support documents, the benefits of basing the summary on the search terms as in THIC is balanced by the benefits of using complete sentences in Abstract summaries.

When we categorized the Live site target documents into subtypes, we found the same pattern of results for all groups. Curiously, the title subtype had the highest accuracy. On inspection, it seemed that these had descriptive titles that were a good match for the tasks. For the 3 other subtypes there are two likely explanations for the similar pattern of results. The findings may be due to the structure of the documents themselves. For example, documents that begin with useless text may be difficult to read, whereas documents that are structured into sections may be easier to read. It is also possible that the similar pattern of accuracy results was due to differences in task difficulty.

Other research [3] has shown the value of grouping search results in categories. For computer support, participants' comments suggest that grouping search results by operating system or hardware platform should make the search process more accurate and faster because users need only look at the applicable category. The benefit of categorizing search results should be greatest when there are a large number of results.

Quality of support and service will become increasingly important in the computer industry as differences among competitive products decrease. We expect to see more efforts to improve customer goal attainment on technical support websites that will further the state of the art in search, summarization, and document authoring tools research.

## 7. REFERENCES


[1] Amitay, E. & Paris, C. (2000). Automatically summarizing Web sites: Is the re a way around it? In *Proceedings of CIKM 2000* (pp. 173-179).

[2] Boguraev B., Neff, M., Lexical Cohesion, Discourse Segmentation and Document Summarization, In *Proceedings of RIAO-2000*.

[3] Dumais, S., Cutrell, E., & Chen, H. (2001). Optimizing search by showing results in context. In *Proceedings of CHI 2001* (pp. 277-284).

[4] Ehrlich K. & Cash, D. Turning Information into Knowledge:Information Finding as a Collaborative Activity. In *Proceedings of Digital Libraries 1994 conference* (pp. 1-8).

[5] Hand, T. & Sundheim, B., "TIPSTER/SUMMAC Summarization Analysis. Tipster Phase III 18-Month Meeting", NIST, Fairfax, Virginia, 1998, Defense Advanced Research Project Agency, Working papers from SUMMAC conference.

[6] Kan, M-Y. & Klavans, J.L., Using Librarian Techniques in Automatic Text Summarization for Information Retrieval, In *Proceedings of the Joint Conference on Digital Libraries - JCDL 2002*; http://www.cs.columbia.edu/~pablo/community/nlp/jcdl02.pdf

[7] Mani, I., Firmin, T., & Sundheim, B., The TIPSTER SUMMAC text summarization evaluation, In *Proceedings of the 9th Conference of the European Chapter of the ACL* (pp.77-58).

[8] D. Radev's Summarization page: http://www.summarization.com/

[9] Summarization Bibliography: http://www.csi.uottawa.ca/tanka/ArtDB/bibliography.html

[10] Summarization resources: http://www.cs.columbia.edu/~hjing/summarization.html


## 8. APPENDIX

**Example "Live Site" Summaries**

[1] PN72528 - OBI MULTIMEDIA MODEM MWAVE PROBLEMS WITH MICROSOFT MAIL (MS MAIL
OBI MULTIMEDIA MODEM MWAVE PROBLEMS WITH MICROSOFT MAIL (MS MAIL

[2] JDK Conversion Assistant
A tool to convert from one Java Development Kit (JDK) to another JDK.

**Corresponding "Abstract" Summaries** (product name obfuscated)

[1] PN72528 - OBI MULTIMEDIA MODEM MWAVE PROBLEMS WITH MICROSOFT MAIL (MS MAIL
[APAR] Typical symptoms: MS Mail sends the dial string to the COM port and then the system seems to lock up. There's never any activity on the phone line. Pressing Ctl-Alt-Del when MS Mail has the focus sometimes allows things to happen, but the system will eventually lock up ...
**Product categories:** Software; Application Infrastructure Services; Networking & Communications; Network File Systems & Sharing

[2] JDK Conversion Assistant
[Downloadable files] The JDK **conversion** assistant helps switch from one **Java** Development Kit (JDK) **Java** 1.2.2 vendor implementation to another by modifying the XXX Server configuration to use the new JDK.
**Product categories:** Software; Application Servers; Distributed Application & Web Servers; XXX Application Server; JDK/SDK; **Operating system(s):** Multi-Platform; **Software version:** 3.0.2 , 3.5

**Corresponding THIC Summaries**

[1] PN72528 - OBI MULTIMEDIA MODEM MWAVE PROBLEMS WITH MICROSOFT MAIL (MS MAIL
Abstract OBI MULTIMEDIA MODEM **MWAVE** PROBLEMS WITH MICROSOFT MAIL (MS MAIL Error Description Typical symptoms: MS Mail sends the dial string to the COM port **...**
**Product categories:** Software; Application Infrastructure Services; Networking & Communications; Network File Systems & Sharing

[2] JDK Conversion Assistant
**…** tool to convert from one **Java** Development Kit (JDK) to another JDK. Download Description The JDK **conversion** assistant helps switch from one **Java** Development Kit (JDK) **Java** 1.2.2 **…**
**Product categories:** Software; Application Servers; Distributed Application & Web Servers; XXX Application Server; JDK/SDK; **Operating system(s):** Multi-Platform; **Software version:** 3.0.2 , 3.5