How Users Assess Web Pages for Information Seeking

Anastasios Tombros
Department of Computer Science, Queen Mary, University of London, London E1 4NS, United Kingdom. E-mail: tassos@dcs.qmul.ac.uk

Ian Ruthven
Department of Computer and Information Sciences, University of Strathclyde, Glasgow G1 1XH, United Kingdom. E-mail: Ian.Ruthven@cis.strath.ac.uk

Joemon M. Jose
Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, United Kingdom. E-mail: jj@dcs.gla.ac.uk

In this article, we investigate the criteria used by online searchers when assessing the relevance of Web pages for information-seeking tasks. Twenty-four participants were given three tasks each, and they indicated the features of Web pages that they used when deciding about the usefulness of the pages in relation to the tasks. These tasks were presented within the context of a simulated work-task situation. We investigated the relative utility of features identified by participants (Web page content, structure, and quality) and how the importance of these features is affected by the type of information-seeking task performed and the stage of the search. The results of this study provide a set of criteria used by searchers to decide about the utility of Web pages for different types of tasks. Such criteria can have implications for the design of systems that use or recommend Web pages.

Introduction

Information retrieval (IR) systems aim to provide users with information that will help them in relation to the information need that they expressed to the system (typically in the form of a query). Searchers are then usually involved in the process of evaluating the utility (or the relevance) of the information (i.e., documents) that the IR system retrieves. One of the most common information-seeking situations entails the use of an Internet search engine (Jansen, Spink, & Saracevic, 2000). The availability of information on the World Wide Web (WWW) has established search engines as a major tool for IR and Web documents as a popular medium through which users access information.

Received August 7, 2003; revised February 12, 2004; accepted February 12, 2004

 ● Published online 15 December 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20106

Assessing the utility of information in relation to an information need is a common task for online searchers. Studies on peoples' perceptions of the relevance of information demonstrate that a range of factors affect human judgements of relevance (e.g., Barry, 1994, 1998; Cool, Belkin, & Kantor, 1993; Maglaughlin & Sonnenwald, 2002; Schamber, 1991). However, such studies often only consider formal textual documents such as journal and conference articles rather than the wide range of formally and informally produced multimedia documents found on the Web. The nature of the IR task on the WWW is different from that on more traditional IR systems (Jansen et al., 2000). One of the differences is the idiosyncrasy of the Web documents themselves. There is generally a large degree of variability in the quality, authority, and layout of Web pages. Moreover, the type of elements such pages contain (e.g., text, multimedia, links) can also vary to a large degree (Woodruff, Aoki, Brewer, Gauthier, & Rowe, 1996), creating a heterogeneous collection of documents distributed over distinct geographic areas.

The motivation behind this study was to gain a better understanding of what features make a Web document useful for information seeking. We concentrated specifically on information-seeking tasks—finding Web pages that contain relevant or useful information—because this is one of the prominent uses of Web pages. It is also a task for which there exist many online tools, for example, search engines (Search Engine Heaven, 2003) and categorization systems (Yahoo, 2003), and many specialized repositories, such as digital libraries (WWW Digital Library, 2003). We observed the decisions made by typical Web users while searching on given information-seeking tasks. We gathered, through think-aloud, questionnaires, system logging, and informal discussion, information on the relative utility of various features of Web documents, such as structural content

(e.g., page layout, link structure), information content (e.g., use of text and multimedia), and aspects of quality (e.g., source of page or recency of information). These features were discovered by users, not predetermined by the experimenters.

Our study concentrated on how Web page features affect people's perception of utility, how the perception changes over the course of a search, and how these issues can be used to influence the design of systems that use Web pages. We also investigated the relationships between these features and the type of information-seeking task given to participants.

The article is organized as follows: We first present some related past work and we outline our aims in the next section, Related Work. We then describe how we structure our investigation: the methodology, subjects used, search tasks given and data collection methods in Methodology. The major contribution of the article is the description and analysis of the results of our investigation. We present the results of the study and a discussion of the results and of their implications in the next two sections. We conclude in the last section, where we also mention some points for taking this work further.

Related Work

Users with information needs expressed in the form of a query submitted to an IR system, such as a Web search engine, may find some information stored in some documents that may be relevant to their needs. Information contained in relevant documents might help the users progress toward satisfying their information needs. Thus relevance is one of the most fundamental, if not *the* fundamental, concept encountered in the theory of IR; and the notion of relevance, whatever that may be, lies at the heart of the IR process.

Despite the fact that the concept of relevance is central to IR, and despite numerous research attempts to precisely define it, a single satisfactory definition has not yet been given (Mizzaro, 1997). Currently, there are two main views of relevance in IR. The first is *topic-appropriateness*, or *topicality*, which is concerned with whether a piece of information is on a subject that has some topical bearing on the information need expressed by the user in the query, and the second is *user-utility*, which deals with the ultimate usefulness of the piece of information to the user who submitted the query (Schamber, 1994). Research into the concept of relevance has indicated that topicality plays a significant role in the determination of relevance (Saracevic, 1970), although topicality does not automatically result in relevance for users (Barry, 1994; Schamber, Eisenberg, & Nilan, 1990). In current IR research, the term *relevance* seems to be used loosely in both senses, despite the fact that the above distinction is widely accepted. In this work, we are mainly concerned with the second notion of relevance, i.e., user-utility.

Mizzaro (1997) has noted that in the past two decades there has been an increase in research efforts to identify user-defined relevance criteria that lead to user-based (or user-utility) relevance. This body of research mainly aims to shed light into the criteria employed by searchers when judging the utility of retrieved information. Several studies have attempted to investigate such criteria (Barry, 1994, 1998; Cool et al., 1993; Maglaughlin & Sonnenwald, 2002; Schamber, 1991; Tang & Solomon, 1998). These studies on peoples' perceptions of relevance of information demonstrate that a wide range of factors affect human judgements of relevance. Some other studies (e.g., Janes, 1991; Saracevic, 1969) have attempted to focus on specific document representations (e.g., titles, abstracts, index terms, full text) and to investigate the effect of these representations on relevance assessments.

Barry (1998), for example, studied factors that influenced searchers' criteria for relevance when reading textual documents. These can be roughly grouped into three main factors:

1. **Content** factors are based on the content of the page or document containing the information. Content factors elicited in Barry's study included the accuracy of the information contained within the document, the availability of the document, and the presence of references to other potentially useful documents.
2. **Personal** factors are based on the relationship between the information and the reader. These include the novelty of the information—how new is the information to the reader—and the reader's ability to understand the information.
3. **Quality** factors are based on the presentation and source of the document. These include the reputation of the document's source, the extent to which the information is clearly presented, and the recency of the document.

Barry's study correlates with findings from other studies, for example, Schamber (1991). Maglaughlin and Sonnenwald (2002) also summarize the findings of this extensive body of research and attempt to identify overlaps between criteria identified in different studies. Maglaughlin and Sonnenwald concluded that for formal textual documents (e.g., journal and conference articles) there seems to be a significant overlap between features identified by various studies, and the field seems to be reaching a consensus regarding what criteria are used in making relevance judgments. Documents found on the Web comprise a wide range of features (Woodruff et al., 1996) (e.g., links and multimedia) and characteristics (e.g., affected by network delays, poor design quality, questionable authority of pages) that differentiate them from the generally authoritative and high-quality research publications involved in these previous studies.

Studies that have specifically focused on features of Web documents that affect searchers' perception of relevance, on the other hand, are few. In a recent study, Kelly, Murdock, Yuan, Croft, and Belkin (2002) examined features of Web documents that influence users' relevance assessments for two types of questions: task-oriented (e.g., How do I get a passport?) and fact-oriented (e.g., How long does it take to

get a passport?). Kelly and her coworkers presented a prespecified range of Web document features to searchers (e.g., lists of items, tables, links, special markups). Their findings showed a difference in the way certain features were used by searchers (e.g., list items occurred more often in documents relevant to task questions, whereas links occurred more often in documents relevant to fact questions). However, although their research focused on Web documents, Kelly et al. considered only a restricted number of predetermined page features instead of allowing searchers to identify themselves the features that allowed them to make the relevance assessments.

Guidelines for assessing the quality of Web pages seek to help Web page designers produce better Web pages. These include pages that are more accessible, readable, or more aesthetically pleasing. Design guidelines range from those composed of high-level design principles, for example, "provide clear visual messages" (Fleming, 1998), to those that are based on empirical investigations of Web page features, for example, word count or use of emphasized text (Ivory, Sinha, & Hearst, 2001). Guidelines for Web design can also help designers of systems that use or recommend Web pages. For example, designers of Web search engines could use Web page quality as an additional scoring factor when retrieving Web pages. In this scenario, higher-quality pages would be recommended before lower-quality pages. However, although there exists a wide range of such guidelines, these guidelines often ignore the tasks for which people use Web pages. That is, the features of Web pages that influence the quality of the page may not be the ones that make the page useful for a given task, or those that are appropriate for different stages along the course of a task.

The fact that people use different relevance criteria at different points in a search has been investigated by, among others, Cothey (2002), Vakkari (2000), Vakkari and Hakala (2000), and Yuan (1997). Search strategies were monitored over a period of time (spanning usually over a number of months) to examine whether searchers adapted their search methods and strategies as they progressed through different stages in their search process. This is an aspect we incorporate into our study, with the difference that the time period we focus on spans the duration of a single search task. We wish to investigate whether searchers' criteria of assessing Web pages change as they advance through different stages in their tasks and become more familiar with the task subject. By asking searchers to describe and explain the features used to assess relevance we can start to prioritize and categorize important indicators of relevance in Web pages. These indicators include ones that can be automatically detected by systems and used to personalize tools to individual users or searchers.

Ingwersen (1992) has identified the analysis of information-seeking tasks and their characteristics as one potential method of improving the effectiveness of IR systems. The effect of information-seeking task characteristics on the search process has been examined by workers such as Kim and Allen (2002); Lazonder, Biemans, and Wopereis (2000); and White and Iivonen (2001). Lazonder et al. (2000) examined the effect of different task types on the search performance of experienced and novice users. White and Iivonen (2001) concentrated on two attributes of tasks (predictability of source and open- or close-ended tasks) and examined how these attributes affect the selection of an initial search strategy (i.e., use of a search engine, direct access of a URL, or use of a Web directory). Kim and Allen (2002) examined the effect of different task types (a known-item search and a subject search) on searchers' activities (e.g., time taken to complete tasks, number of Web documents viewed, number of search iterations) and search effectiveness. Dependence on task characteristics was discovered in most of these cases. For example, for known-item searches, search effectiveness was higher and searchers spent less time to complete the task than for subject searches. In our study, we examined how different task types affected the features that searchers employ to judge the utility of Web pages. This aspect of task dependence is not widely explored in the literature.

The major contributions of this work are therefore threefold. First, we identified the Web document features that searchers employ to make assessments of the utility of documents in relation to information seeking tasks. As we discussed previously, features of Web pages that affect utility assessments have not been widely investigated, with the exception of (Kelly et al., 2002). In contrast to Kelly et al. however, we did not present searchers with a predefined set of features. Instead, we discovered the set of features that searchers themselves employ during the tasks. Second, we examined the effect of different types of information seeking tasks on the criteria employed by searchers, as well as on the effectiveness of the searches. Third, we looked into how searchers' criteria evolved during the different stages of tasks. In contrast to the majority of other work in this area, we focused on the span of a single search session (rather than on a search task spanning over a long period of time). This type of examination is better tailored to search scenarios on the WWW.

Our study, therefore, has important implications for the design of systems that recommend or use Web pages. A good example is Web page summarization systems, for example, White, Jose, and Ruthven (2003). These systems offer searchers short summaries of a retrieved Web page, often consisting of a number of sentences taken from the content of the page. Although these summaries have been shown to be useful for interactive searching, they generally only consider the textual content of the page. By asking what textual *and* nontextual aspects of pages users employ, systems can provide a more complete picture of Web pages. By considering more information on the task itself, summarization systems could concentrate on selecting the most useful aspects of the Web page. These useful aspects could come from the whole range of textual, nontextual, and quality features of pages. In this way it may be possible to allow searchers to more effectively detect the "information scent" (Chi, Pirolli, Chen, & Pitkow, 2001) of Web documents.

Methodology

The study involved 24 participants, each of whom was asked to search on three information-seeking tasks. We allowed the participants to search for information in any way they found useful or natural, encouraging them to search as they would normally. Participants could use any publicly available Web search engine, personal knowledge of the Web (e.g., useful URLs or Web sites), or search tools. Therefore, there was no restriction placed on the initial search strategy that participants would employ (White & Iivonen, 2001). The only restriction placed was that they could not ask for further information on the topic while searching; for example, they could not ask the experimenter for recommendations on good places to search or strategies to use. This was mainly to ensure that participants performed the tasks in as unbiased a setting as possible. One of the experimenters (Tombros) was present in the laboratory during the study.

Participants were asked to discuss, in the form of think-alouds, their perceptions of what constituted useful information on the Web pages they chose to view. All participant utterances were recorded, along with the desktop activity, by using the *Camtasia* software tool (Camtasia is a product of Techsmith corporation, http://www.techsmith.com). The experimenter present in the laboratory used at times a neutral questioning technique to probe answers from searchers in cases where they were reluctant to do so. The outcome of this logging was a video clip for each task that each participant attempted (total of 72 clips). The recorded video clips were subsequently examined by the experimenters (for more details, see the section on Extracting Information From Search Sessions). After each search task, participants were asked to complete a questionnaire that asked for their opinions on the task, such as whether they felt they had completed the task, and information on the features of the Web pages they viewed. In addition, each participant completed a questionnaire before the start of the study (see Questionnaires section). We provide the instructions that were handed to searchers in the Appendix at the end of this article.

Participants

Twenty-four participants (17 male and 7 female) were recruited for the study. They were recruited by word of mouth and personal contact through e-mail messages posted to student mailing lists. Participants received payment for their participation in the study. The 24 participants were split in two groups of 12 participants each. One group was given 15 minutes to search on each task, while the second group were given 30 minutes per task, a time we felt was sufficient for participants to complete each task. We initially allowed 15 minutes per task, but we noticed that most of the first 12 participants did not manage to proceed sufficiently in all their tasks. We therefore decided to increase the allowed time per task for the remaining 12 participants. This would also allow us to compare the set of criteria used by the two groups and examine whether time pressure had an effect on the page features employed by participants.

The age of the participants ranged from 20 to 41, with an average of 27. Most of them were affiliated to the University of Glasgow, either as undergraduate or postgraduate students, or as academic staff (lecturers, research assistants, and administrative staff). Eleven of the participants were affiliated with the department of computing science; six, with engineering departments; two, with social sciences; and one, with statistics. All participants had to complete a questionnaire prior to the start of the study. The questions aimed at collecting information about the participants' familiarity with searching the Web for information and their perception of how successful their searches typically are. The majority of the participants browse the Internet and interact with search engines several times a day (16 and 15 out of the 24 participants, respectively) and use Google for their searches (19 out of 24 participants). Also, most of the participants tend not to use advanced features of search engines (13 participants) and tend to make use of online bibliographic databases (16 participants). As far as their perception of the success of their online searching is concerned, most of the participants think they often find what they are looking for on the Web.

Search Tasks

Each of the participants was asked to complete three search tasks. The search tasks were given rather than using tasks from analyzing real searches, to allow a comparison between different people searching on the same tasks. The search tasks were in the form of short search scenarios such as the one shown in Figure 1. These scenarios were intended to provide the searcher with background motivation to the search and sufficient contextual detail to decide on the relevance of viewed Web pages. This is in contrast to simply

> You are considering a career as a web-page designer and have an interview next week with a company you really want to work for. The position will involve designing sites to allow local companies to sell their products on-line. You realize, however, that you know little about who actually uses the Internet.
>
> To impress your future employers you think it is a good idea to get some information on what kind of people have Internet access so you can discuss how you would design sites to attract these groups of people.

FIG. 1. Search scenario.

asking the searchers to find specific pieces of information, for example, "Find demographic information on users of the Internet."

The search scenarios did not outline what specific information was required to complete the search task. Rather, the decision on what constituted relevant information, and whether or not the search task had been completed, was made by the searcher. The use of *simulated* search situations such as these encourage users to treat the information-seeking task as a personal task, searching as though the task was their own (Borlund, 2000).

In our study we used three search tasks. These tasks not only asked the subjects to find different information but simulated different types of search task.

- **Background search.** In this task the participants were asked to find general background information on a topic, essentially as much information as possible on a topic. In our study the participants were asked to find information on the demographics of the Internet.
- **Decision task.** In this task the participants had to gather information and make a decision based on the information found while searching. The participants, in this case, were asked to decide on the best hi-fi speakers available in their own price range.
- **Many items task.** In this task the participants were asked to compile a list of items. This task specifically asked the participants to compile a list of interesting things to do over a weekend in the city of Kyoto.

The order of presentation of the tasks was rotated among participants to avoid any learning effects.

Questionnaires

Apart from the questionnaire before the start of the study, each participant had to complete a questionnaire before (presearch) and after (postsearch) each task, as well as a final questionnaire after all tasks. The presearch questionnaire was aimed at measuring the participants' familiarity with the task topic, as well their expectation of how easy it would be to find the necessary information. Likert scales (Preece, 1994) were used in designing the presearch questionnaire.

The postsearch questionnaire was aimed at collecting data on three issues: (1) to measure the participants' perception of task completion; (2) to examine whether participants found the task topic clear, easy, interesting, familiar, relevant to themselves, and stressful, as well as to examine whether their searching behavior during the task was similar to their normal searching behavior; (3) to assess the participants' perception of importance of certain aspects of the Web pages they viewed (e.g., text, multimedia, layout, knowledge of the topic) in terms of helping them to determine the usefulness of pages. This issue should not be confused with collecting information on the importance of features through log analysis during the three tasks (see later in the section on Extracting Information From Search Sessions). Questions in the postsearch questionnaire were formulated using Likert scales and semantic differentials. The final questionnaire asked participants to rate the tasks based on their difficulty, as well as to provide any further comments they felt necessary. The questionnaires used in this study are given in the Appendix.

Extracting Information From Search Sessions

Information regarding the importance of features was extracted through the analysis of the recorded search sessions. Features were identified through the reasons mentioned by participants for assessing a Web document as useful or not useful. Participants explicitly stated how they characterized each document they visited (e.g., "I would find this page useful because . . ."). Similar reasons were grouped together to form a single criterion (Flick, 1998). We analyzed each recorded user session (i.e., the user's interaction with the Web browser, as well as the user's speech explaining their judgments and the features they mention) and extracted Web page features that were mentioned by participants as indicating a page's usefulness (*positive mention*) or nonusefulness (*negative mention*) to a task (i.e., features were discovered through participants' remarks rather than predefined by the experimenters). In some cases, participants mentioned the lack of a feature (e.g., if this page showed a picture, it would be useful). Such responses were coded as positive mentions for the features.

We thus gathered data for features in two lists, depending on the type of pages (useful or not) in which they appeared (Cool et al., 1993). It should be mentioned that both positive and negative mentions of features should be seen as an indication of the feature's importance. Regardless of whether the feature was used to indicate a useful or a not-useful page to a user, the mention of a feature is treated as evidence of its importance. The features were also grouped under broader categories (e.g., text, structure, quality). For example, the *structure* category comprises the features layout, links, links quality, and table layout. The set of all features and categories identified through this process is presented in detail in the section on Overall Importance of Features.

Only one of the authors (Tombros) was involved in analyzing the search sessions. To counterbalance for the lack of cross-validation of the assigned features, the same experimenter reanalyzed 24 of the 72 search sessions two months after the study was completed. The agreement between the two sets of features was 95%. The initial set of features was used in the study. We feel that in the lack of intercoder agreement data, this high agreement provides evidence for the reliability of the discovered document features.

In Table 1 we present the number of Web documents judged by the participants. We present separate results for documents that were judged as useful (columns 2–3) and for those judged as not-useful (columns 4–5) for both the 15- and 30-minute groups. In column 6 we present the total number of documents judged.

TABLE 1. Number of Web documents judged.

	Useful		Not useful		
	15'	30'	15'	30'	Totals
Task 1	37	74	90	115	316
Task 2	40	65	76	76	257
Task 3	56	74	91	68	289
Totals	133	213	257	259	862

Results

In this section we present the results obtained from the analysis of the user sessions. First, in the section on Categories and Features we present the features and categories of features that were discovered during the study. Next, in Overall Importance of Features, we look into the overall importance of document features across all tasks and regardless of positive or negative mentions. Then, in Positive Versus Negative Mention of Features, we examine any differences between positive and negative mentions of features, and in Variation of Feature Importance Across Search Tasks, we examine the variation of feature importance across tasks. In Time Constraints: 15-Minute Versus 30-Minute Group, we present results about any differences in features used based on the time limits that users faced, and in Progression of Criteria Along Tasks, we examine the way that the features used by participants evolved during their progress along the tasks.

Categories and Features

In this section we present in detail the various features and categories that were identified during the study. The features and categories are shown in Table 2.

Text. Features in this category capture various textual aspects of a Web document. Such aspects include the general content of the document (content), numerical figures in the document (e.g., dates, currency data, numbers), content of the document that contained some of the user's query terms (query terms), and content of the document that is located in the title or section headings of the document (title/headings). The extent to which some Web documents contain an overwhelming amount of text is also captured in this category (too much text).

Structure. Under the structure category, we include features that pertain to structural aspects of a Web document. The general layout of the page (layout) refers to the general format of a Web document and the way information is presented in it. The links contained in a Web page (links) are also included in this category, together with the presence of any tabular data in the document (table layout). The feature link quality refers mainly to cases where participants were overwhelmed by the number of links present in a Web page.

TABLE 2. Number of mentions of document features.

	Useful		Not useful		Combined	
	#	%	#	%	#	%
Text	**367**	**46.69**	**349**	**42.77**	**716**	**44.69**
Content	185	23.53	204	25	389	24.28
Numbers	109	13.87	49	6	158	9.86
Titles/headings	37	4.71	34	4.17	71	4.43
Query terms	34	4.33	29	3.55	63	3.93
Too much	2	0.25	33	4.04	35	2.18
Structure	**176**	**22.39**	**170**	**20.83**	**346**	**21.60**
Layout	60	7.63	95	11.64	155	9.68
Links	80	10.18	28	3.43	108	6.74
Links quality	5	0.64	37	4.53	42	2.62
Table data/table layout	31	3.94	10	1.23	41	2.56
Quality	**133**	**16.92**	**150**	**18.38**	**283**	**17.67**
Scope/depth	28	3.56	59	7.23	87	5.43
Authority/source	61	7.76	23	2.82	84	5.24
Recency	31	3.94	35	4.29	66	4.12
General quality	8	1.02	25	3.06	33	2.06
Content novelty	5	0.64	4	0.49	9	0.56
Error on the page	0	0	4	0.49	4	0.25
Non-textual items	**99**	**12.60**	**44**	**5.39**	**143**	**8.93**
Pictures	99	12.60	44	5.39	143	8.93
Physical properties	**11**	**1.40**	**103**	**12.62**	**114**	**7.12**
Page not found	0	0	36	4.41	36	2.25
Page location	6	0.76	16	1.96	22	1.37
Page already seen	1	0.13	16	1.96	17	1.06
Language	1	0.13	4	0.49	5	0.31
File type	0	0	4	0.49	4	0.25
File size	2	0.25	1	0.12	3	0.19
Connection speed	1	0.13	13	1.59	14	0.87
Subscription/registration	0	0	13	1.59	13	0.81
Totals	**786**		**816**		**1602**	

Quality. This category is rather wide, in that it encompasses a number of features referring to qualitative aspects of a Web document. Such features include the scope and depth of the information contained in the document (scope/depth); the authority of the source of information contained in the document (authority/source); the recency of the information (recency); the overall quality of the Web page in terms of appearance, formatting, and the like (general quality); the novelty of the information contained in the page (content novelty); and the presence of any actual errors (such as HTML errors) on the page.

Nontextual items. Information items that are of a nontextual form. In the context of the tasks performed, our users only came across pictures (i.e., no video or sound items were encountered), and therefore the only feature in this category corresponds to pictures.

Physical properties. This category comprises features that pertain to physical characteristics of Web documents: the size of a Web document (file size), the speed with which it is downloaded (connection speed), the actual geographical

location of a document (page location as identified through the URL and the content of the page), whether a page has been previously seen by a user, whether a page cannot be found, the file type corresponding to a Web document (e.g., acrobat reader files, html documents), the language of the document (i.e., English or otherwise), and finally whether one needs to register or subscribe to an online service to access the document. Of these five categories, the quality category can be seen as being more subjective than the other four. One could argue, for example, that the features in the text and nontextual categories are objective, corresponding to features found in Web pages: Any observer could see whether a page contains useful currency data or multimedia items. Some of the features contained in the quality category, however, measure more subjective qualities of the information contained in Web pages (e.g., general quality, content novelty). Some of these subjective qualities are also based on page features found in other categories. For example, the recency of a page can be inferred by looking at the content of the page (category: text; feature: content). Previous studies of relevance (e.g., Barry, 1994) have differentiated between such features and information qualities.

However, for the purpose of this study, we treat all five categories as equivalent. This is because our aim is to identify which aspects of Web pages (whether subjectively or objectively assessed) searchers employ when making utility judgments. We specifically attempt to identify the implications these aspects may have for the design of systems that use or recommend Web pages. With this aim in mind, we view all mentions made by searchers as representing aspects of Web pages that were used when making utility assessments. The features collected under the quality category should therefore be seen as contributing toward this direction. For example, it may be the case that some features from the text category were employed by searchers when mentioning the scope/depth feature. However, the fact that the features were mentioned by searchers with the specific aim of identifying the scope and depth of the information in Web pages as contributing to the utility assessment, makes this objective feature (scope/depth) important for consideration. In other words, we view the categories and features quantitatively and try to extract as many instances where features are mentioned as possible.

Overall Importance of Features

In this section we present results about the overall importance of features. The data for all three tasks are presented in Table 2. The features and categories are reported in the first column of the table, and the data for each feature and category are presented in columns 2–7. Data are presented separately depending on whether features were mentioned to indicate that a Web page was useful to a task (columns 2–3), not useful to a task (columns 4–5), and overall (i.e., the sum of the previous two, columns 6–7). More specifically, in columns 2–3 we present the absolute number of mentions and the percentage with respect to the total number of mentions, respectively, that each feature and category acquired when mentioned in relation to a Web page that users found useful to their task (*positive mentions of features*). The total number of positive mentions is recorded at the bottom of column 2. In columns 4 and 5 we present similar data for the cases where features were used to indicate Web pages that were not useful to the users' tasks (*negative mentions of features*); and in columns 6 and 7, the total data that result from the addition of positive and negative mentions.

The results in this section are based on the *combined* columns of Table 2 (i.e., columns 6 and 7). Based on these figures, we can see that text is the most important category in determining the usefulness of Web documents for online searchers. A feature from the text category was mentioned in almost half of the total feature mentions. The two most important individual features (content and numbers) also belong to this category. The more generic feature content displays the largest number of mentions in the category text. A contributing factor to this is that sometimes it was not possible to extract a more refined account of what textual features searchers employed. In other words, some of the mentions attributed to content may actually account for mentions of other textual features (see also the section on Limitations).

Structure is the second most important category. Two of the most important individual features mentioned by users (layout and links) belong to this category. Quality follows as the third most important category, with three of its features (scope/depth, authority/source and recency) displaying a relatively high importance among all features. The existence of only one type of nontextual information (pictures) in the pages visited by the participants means that the respective category scored rather low in the overall importance of categories (fourth out of five). The only feature in this category however, was the fourth most important feature overall. The least important category of those encountered in our experimental setting was the physical properties of Web documents. It should, however, be mentioned that there is a great imbalance between the negative and positive mentions of features belonging to this category (as is obvious from Table 2, negative mentions are considerably more than positive mentions for this category). This imbalance stems from the nature of the features assigned to this category and will be further discussed in the section on Positive Versus Negative Mention of Features.

As we mentioned previously, the participants assessed the utility of 862 Web pages by providing 1,602 mentions of page features, giving an average of almost two (1.9) features per assessed page. There were also many cases where participants utilized only a single feature to indicate a page's usefulness or nonusefulness (371 cases in total, 78 positive mentions vs. 293 negative mentions). The results we have so far presented in this section only refer to overall mentions of features, with no consideration to the patterns of occurrence of features (i.e., do certain features tend to co-occur with certain features, does the presence of an individual feature warrant a strong indication of usefulness or nonusefulness of a document). We examine such issues in the next two sections.

TABLE 3. Number of times a feature was mentioned as a single indicator of usefulness (total of 371 single mentions).

Content	111 (28.7%)	Links quality	16 (38.1%)
Query terms	24 (38.1%)	Links	15 (13.9%)
Scope/depth	21 (24.1%)	Pictures	12 (8.4%)
Layout	17 (11%)	Authority/source	10 (12%)
Recency	17 (25.8%)		

Single strong indicators. We first examine the features that were used as single indicators of usefulness or nonusefulness of Web pages. Such features can be seen as strong indicators of Web document utility. In an operational environment, such strong indicators can serve as effective document representations that would inform users of the potential utility of Web documents.

As mentioned before, for 371 out of the 862 documents judged, a single feature was provided as an indicator of the document's usefulness to a task, and in their vast majority such single mentions were used to indicate nonusefulness. In Table 3 we present the number of mentions (both positive and negative) of the most important features as single indicators and, in brackets, the percentage over the total mentions for these features that this number represents. For scope/depth, for example, 24.1% (21 mentions, Table 3) of its total mentions (87 total mentions, Table 2) were made as single mentions of usefulness/nonusefulness. The two most mentioned single indicators of usefulness belong to the text category (content and query terms). Features that belong to the structure and quality categories follow in the ranking, which in general seems to reflect the ranking of features obtained from Table 2.

The number of mentions of features as single indicators of usefulness can be seen as further strengthening the results presented in Table 2. Features such as text, query-term specific text, scope/depth, and layout are not only important when considering the overall number of mentions but are also important as strong indicators of document usefulness. In practical terms, the implication of these results is that if these features of a Web document are captured and presented to the user as a preview of the document, then the user will be more likely to make an accurate prediction of the document's usefulness without needing to refer to its full text.

Co-occurrence of features. In this section we examine patterns of co-occurrence of document features. We aim to examine whether certain features tend to be mentioned together when searchers decide on the usefulness of Web documents or, in other words, whether the presence of one feature predicts the presence of another feature. Such relationships are not revealed by the data in Table 2. Because the data gathered are of a binary form (i.e., presence or absence of a feature mention for each assessed document), an appropriate statistical method is binary logistic regression (Hosmer & Lemeshow, 1989). This method estimates the likelihood that a response variable (i.e., one of the features) can be predicted on the basis of some of the other variables (i.e., the rest of the features). Only significant relationships between variables are considered (i.e., those for which the probability that the relationship is random is less that 0.05).

The strong presence of the content feature means that it co-occurs with a large number of other features but not strongly enough to assume that its presence either predicts, or is predicted by, the presence of other features. The only exception was noted for the first task (background search), where the presence of the feature "table data" predicted the presence of content ($p = 0.02$). As far as the other features are concerned, a few significant relationships were discovered. For the first task, numbers and table data predicted the presence of each other ($p = 0.007$), whereas for the second task (decision search) it was pictures and authority/source that significantly co-occurred ($p = 0.02$). For the third task (many items search), it was numerical data and links ($p = 0.005$), as well as pictures and links ($p = 0.02$) that significantly co-occurred.

These results suggest that there are small "clusters" of important features for individual tasks. The importance of these features is identified by their pattern of co-occurrence for the tasks. For example, for the background search, numerical data and table data co-occur significantly enough so that the presence of one predicts the presence of the other. Participants seem to jointly mention these two features when assessing pages for this task, because they seem to be important for gathering information specific to the task (demographics of Internet users). One can therefore argue that these two features correspond to important aspects of pages in relation to the first task. The same applies to the features discovered for the second and third tasks.

Positive Versus Negative Mention of Features

By observing the data in Table 2 we can see that the importance of document features changes depending on whether searchers are judging pages as useful or not useful to their tasks. In the previous section we discussed the relative importance of features with no respect to whether they acted as indicators of useful or not useful Web pages (i.e., positive or negative mentions, respectively). In this section we examine whether there is a difference depending on the type of pages indicated.

Features that exhibit the tendency to mainly indicate either useful or nonuseful pages are equally important, because they both inform searchers whether a certain page is worth their time and effort. Features with predominantly positive mentions can inform users of the potential information value in the document, whereas features with mainly negative mentions may reduce number of false hits, i.e., situations where users may visit a page only to discover that it is not useful to their information need.

Participants in general mentioned approximately the same number of features for pages judged as useful and not useful (786 vs. 816, respectively, Table 2), but judged more pages as not useful than as useful (516 vs. 346, Table 1). The small difference in the number of feature mentions despite

TABLE 4. Feature ranking for useful and not useful documents.

Useful	Not useful
Content (185)	Content (204)
Numbers (109)	Layout (95)
Pictures (99)	Scope/depth (59)
Links (80)	Numbers (49)
Authority/source (61)	Pictures (44)
Layout (60)	Links quality (37)
Titles/headings (37)	Page not found (36)
Query terms (34)	Recency (35)
Recency (31)	Titles/headings (34)
Table data (31)	Too much text (33)

the large difference in the number of pages judged as useful or not useful can be explained by taking into account that for 293 pages judged as not useful participants mentioned only a single feature as supporting their assessment. When judging pages as useful, though, participants tended to be more elaborate in their judgments. This behavior seems to be in agreement with findings by Maglaughlin and Sonnenwald (2002), who suggest that participants may examine useful documents more carefully or find it easier to discuss positive associations between their information needs and documents.

In Table 4 we present the overall ranking of the 10 most important features depending on whether we take into account the positive (first column of the table) or negative (second column) mentions of features. The data in this table demonstrate that content remains the most important feature regardless of the type of pages indicated. A notable change occurs for layout, which becomes the second most important indicator of nonusefulness (compared to being the sixth most important indicator of usefulness). Moreover, links and authority/source, which are both important features for determining useful pages, do not appear in the top 10 features for determining not-useful pages. Scope/depth similarly appears to be a highly important indicator of nonusefulness but does not appear to be significant for determining useful pages.

In Table 5 we present further data on the use of features for positive and negative mentions. More specifically, the left column of the table lists the features with more positive mentions. Next to each feature is the percentage of the total mentions of this feature that are positive and, in brackets, the difference between the number of positive and negative mentions. In the right column of the table we provide similar data for features with more negative mentions (a negative number for a feature indicates more negative than positive mentions). For example, 61% of the total mentions of the feature layout are negative, translating into 35 more negative mentions.

TABLE 5. Difference between positive and negative mentions for the most important features.

Table data	76% (21)	Links quality	88% (−32)
Links	74% (52)	General quality	76% (−17)
Authority/source	73% (38)	Page location	73% (−10)
Numbers	69% (60)	Scope/depth	68% (−31)
Pictures	69% (55)	Layout	61% (−35)
Query terms	54% (5)	Content	52% (−19)

Translating the different mentions for individual features into different mentions for feature categories, we note that physical properties display 92 more negative mentions; nontextual items (i.e., only the feature pictures), 55 more positive mentions; text, 18 more positive mentions; quality, 17 more negative mentions; and, finally, structure, six more positive mentions.

The significantly increased number of negative mentions for the features belonging to the physical properties category can be seen as a consequence of the nature of these features. Some of the features assigned to this category can only be mentioned in relation to a page that was not useful for the user's task (e.g., when a page cannot be found on the server, or when a user needs to register or subscribe to an online service to gain access to information). The majority of the features in this category, however, could either have a negative or a positive mention (e.g., the geographical location of a Web page, the file type, and file size). Despite this, participants employed such features, almost always, only for negative mentions (11 positive mentions vs. 103 negative).

It is also worth noting that the mentions for the text category are balanced between positive and negative. This is not only true for the overall mentions of the category, but it also applies to the individual features within the category (e.g., content, query terms, and titles/headings all have balanced positive and negative mentions).

Variation of Feature Importance Across Search Tasks

In this section we examine whether there is a difference in the number of mentions of document features across the three tasks used in the study. Such differences may exist because the nature of a specific task may constitute certain document features as more important than others at judging usefulness. For example, in the second task we ask searchers to find a pair of hi-fi speakers to fit their budget. One may therefore expect that visual information (i.e., pictures) is to be of higher importance for this specific task than for the first task, where we ask participants to locate demographics of Internet users.

In Figure 2 we present the mentions of the most significant features (according to Table 2) across the tasks. In this figure the average percentage of mentions for each feature is plotted for each of the three tasks. We can notice that some features are considerably biased toward (or against) specific tasks. Such examples are the limited use of pictures for task 1, the increased use of numbers for task 2, and the increased use of links for tasks 1 and 2. In Table 6 we present the average number of mentions for each feature across all three tasks and the standard deviation of the mentions.

From the data in Figure 2 and Table 6 we can observe that mentions for content, layout, and authority/source are relatively evenly distributed across the three tasks. Some other

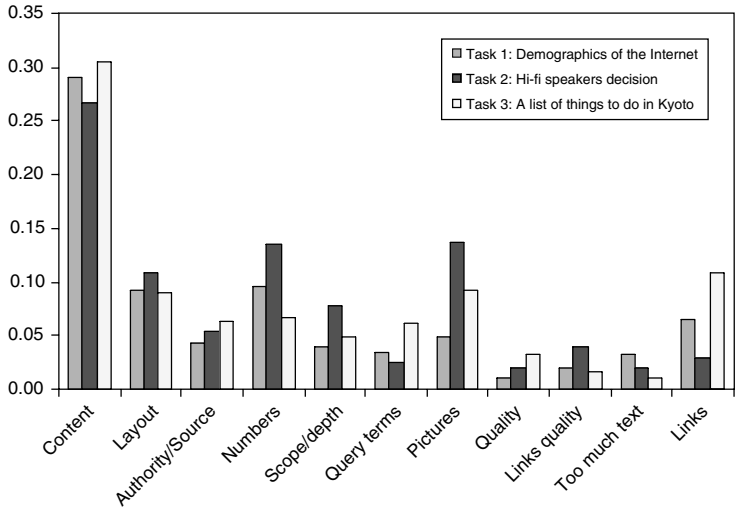

FIG. 2. Feature mentions across tasks.

features (e.g., pictures, numbers) display much higher mentions for specific tasks (i.e., task 2 for pictures and numbers). This is attributed to the specific characteristics of task 2 (decision task), in which users need to rely on visual clues (i.e., pictures of hi-fi speakers) and numerical characteristics (i.e., prices, specifications) to accomplish the task. It should be mentioned that this task involves users in a great deal of comparisons between the items they are considering. Factual aspects of documents in this case, either textual or visual, become salient in helping users assess the utility of pages.

Another interesting result is regarding the use of links. Mentions for this feature are heavily biased mainly toward task 3, which is a many-items task, and partly toward task 1, which is a background task. For a many-items task, searchers seem to make more frequent use of links to pages of potential related interest to access pages (e.g., pages with information about tourist attractions, events, exhibitions, concerts) that would help them compile the many-items list required by the task. For the background type of task 1, links offer the opportunity to explore the topic at hand and to be able to gain access to additional sources of information that would help searchers gain a better understanding of the topic. It should also be noted that the increased mention of the links quality feature for task 2 (Figure 2) refers mainly to searchers assessing pages as not useful because of the large number of links they contained (i.e., links pages). For this decision type task, searchers required more factual features than links to other documents of related topics.

TABLE 6. Average and standard deviation of feature mentions across tasks.

Features	Average	Standard deviation
Content	0.287	0.019
Numbers	0.099	0.035
Layout	0.097	0.010
Pictures	0.093	0.044
Links	0.068	0.040
Scope/depth	0.056	0.020
Authority/source	0.053	0.009
Query terms	0.040	0.019
Links quality	0.025	0.012
Quality	0.021	0.010
Too much text	0.021	0.011

Some further differences are evident for the use of query terms (increased mentions for task 3) and the use of scope/depth (increased mentions for task 2). Query terms in task 3 were mainly used in pages with long textual contents to filter out unnecessary information and focus on the items of interest. The specific nature of this task encouraged this type of behavior. For example, a participant interested in music events in Kyoto would often use the search function of the Web browser to look for occurrences of "music" in a page. As far as the scope/depth feature is concerned, its increased use in a decision task is based on that users required enough information in pages (e.g., enough details about prices, specifications, guarantees, availability of speakers) to make an informed comparison of the available choices.

We also collected data regarding the participants' perception of the three tasks by means of postsearch questionnaires and a ranking of the tasks' difficulty by means of the final questionnaire (see the Questionnaire section). Based on the final questionnaire, users judged the first task as the most difficult (average ranking of 1.4, 1 being the most difficult and 3 the least difficult), followed by the second task (average ranking 2.1) and the third (average 2.4). It should be mentioned that 16 out of the 24 participants rated task 1 as the most difficult, 5 rated task 2 as the most difficult and 3 participants thought task 3 was the most difficult.

The ranking of the tasks' difficulty was reflected in the participants' perception of the three tasks. We measured participants' perceptions in the postsearch questionnaire, on a 5-point Likert scale, where a mark closer to 1 corresponds to a stronger agreement with a statement. Testing for significance was done using the Wilcoxon signed-ranks test. For

the most difficult task (task 1), users had a lower satisfaction with the outcome of the task and a lower task completion perception than for either of the other two tasks ($p = 0.05$ against task 2). Moreover, participants found this task less clear (1.9 vs. 1.3 and 1.3), more complex (2.7 vs. 3.7 and 3.3, $p < 0.001$ against task 2 and $p = 0.01$ against task 3), less familiar (2.3 vs. 3.7 and 4.1, $p < 0.001$ against both tasks), more stressful (3.4 vs. 4.1 and 3.8), less relevant to themselves (2.4 vs. 3.1 and 3.5, $p = 0.003$ against task 3) and less interesting (3.2 vs. 3.5 and 3.8) than tasks 2 and 3, respectively. In addition to the questionnaire data, participants took on average longer to complete task 1 (average time 18.3 minutes) than task 2 (16.3 minutes) and task 3 (15.6 minutes). The differences between task 1 and 3 are significant ($p = 0.03$).

Time Constraints: 15-Minute Versus 30-Minute Group

In this section we examine variations in the criteria used between the two groups of users depending on the time constraints they faced. In the section on Methodology we mentioned that half of the users had 15 minutes to complete each task, while the other half had 30 minutes. Although the comparison between the two groups was not a primary focus of our study, the information collected lends itself to examination. It should also be mentioned that because two different groups of participants formed the two groups, a strict comparison of the data in this section is not attempted. Instead, we aim to get an overview of the trend of feature mentions between the two groups. In Table 7 we present the percentage of mentions for each of the most important features mentioned by users in the two timed groups. In general, there do not seem to be large differences in the way the two groups of users employed page features. The most notable differences occur for the increased use of numbers from the 30-minute group, on the one hand, and the increased use of links and links quality, scope/depth, and query terms from the 15-minute group, on the other hand.

One interpretation of these differences is that participants in the 15-minute group did more filtering of the Web documents by relying on more "obvious" features than those in the 30-minute group. Features such as query terms, links quality (in terms of the number of links present in a Web page), and links can be seen as more obvious features that do not require an in-depth examination of the content and structure of a Web document. The query terms feature, for example, was often rapidly employed by users through the Web browser's search function to locate the presence of query terms in a document. Participants in the 15-minute group might have used such features to compensate for the less time made available to them. Participants in the 30-minute group, on the other hand, had more time to employ features that required a more careful examination of the documents (e.g., content, numbers).

The most significant difference between the two groups of participants was the higher perception of task completion for participants in the 30-minute group. More specifically, users in the 15-minute group had a lower perception of task completion and outcome satisfaction than those in the 30-minute group for all three tasks. On the 5-point Likert scales the differences were 3.4 vs. 2.6 for task 1, 2.5 vs. 2 for task 2, and 3.4 vs. 1.6 for task 3 (averages for the 15-minute group are given first, lower values represent higher perception of task completion). In addition, searchers in the 15-minute group found the three tasks more stressful and complex than those in the 30-minute group. This is also reflected in the average time taken by the two groups of users to complete tasks. The 15-minute group took on average (for all three tasks) 94% of the allowed time (i.e., 14.1 minutes average per task), while the 30-minute group took 64.7% of the allowed time (i.e., 19.5 minutes average per task). It is worth mentioning that for the first task, 11 out of the 12 searchers in the 15-minute group used all of the allowed time (this is in contrast to only 2 of the 12 participants in the 30-minute group).

Progression of Criteria Along Tasks

To examine if there is a variation in the document features participants employ along the duration of a task, we needed a method of identifying various stages in the participants' progress through a task. The methodology we employed is as follows. For each task a participant attempted, we identified the first and last set of Web documents the participant visited. The sets were identified in relation to the first and last query submitted to a search engine, or to the first and last direct access to a Web site. We assume that the set of features employed in assessing the first and last set of Web documents viewed will be representative of any shift of focus that might have occurred in the employed features along the duration of a task. It is worth noting that there is an underlying assumption in this methodology. We assume that the pattern of mentions between the first and last set of documents examined is representative of the pattern of mentions in other points of the task that lie in between. We chose to follow this specific methodology for practical purposes, because it was easier to capture feature mentions at these two points in the task. In Table 8 we present, for each

TABLE 7. Overall mentions of features for the 15- and 30-minute groups.

	Content	Layout	Numbers	Pictures	Links	Scope depth	Query terms	Recency	Authority source	Links quality	Too much text
15 min.	27.37%	10.26%	8.24%	8.24%	8.09%	6.53%	5.60%	4.98%	4.35%	4.04%	2.80%
30 min.	29.77%	9.33%	11.01%	9.43%	5.87%	4.72%	2.83%	3.56%	5.87%	1.36%	1.78%

TABLE 8. Variation in feature mentions between first and last query in a task.

		Layout	Numbers	Links quality	Content	Query terms	Pictures	Scope/ depth	Authority	Links	Quality
Task 1	First	9.4%	8.6%	2.3%	31.2%	3%	4.5%	4.5%	2.6%	5.6%	1.1%
	Last	8.1%	9.3%	3%	26.7%	4.2%	5.5%	3%	4.2%	7.2%	1.3%
Task 2	First	9.3%	16%	4%	29.3%	2%	14%	4%	5.3%	0.4%	0.3%
	Last	13.8%	12.2%	2.7%	26.4%	3.4%	14.9%	8.1%	6.8%	2.7%	3.4%
Task 3	First	11.3%	8.4%	2.1%	29.9%	6.9%	7.4%	3.4%	5.4%	11.8%	2.9%
	Last	4.5%	6.1%	0%	36.5%	5.1%	13.2%	5.9%	4.5%	10.3%	1.9%

task, the average percentage of mentions for features for the first and last set of documents examined by participants. From the data presented in Table 8 we notice that there is a certain degree of variation in the criteria participants employ near the start and near the end of a task.

For the first task (background task), the variation in the features is generally smaller compared to the other two tasks. This may be explained by the difficulty of this task, as perceived by participants (see the section on Variation of Feature Importance Across Search Tasks). It is likely that because of the difficulty of this task, participants did not have the opportunity to adjust their criteria of assessing page utility. Participants seem to employ more obvious criteria at the end of their task compared to the start. For example, content mentions seem to decrease at the end of the task, whereas query terms, pictures (mainly graphs summarizing statistical information), and the authority of the Web pages seem to increase. This behavior seems to suggest that participants adjust their strategy to look into more so-called superficial features of documents as they progress along the task. Given the difficulty searchers had in finding useful information for the first task (see Variation of Feature Importance Across Search Tasks, earlier), these more superficial features allowed them to more quickly and easily filter out irrelevant documents.

For the other two tasks there are some more pronounced differences in feature mentions. For the second task (a decision task), participants seem to progress through a stage of high mentions of content, numbers (especially prices and technical specifications), and links quality (in terms of too many available links in the page) at the start of the task, to high mentions of page layout, scope and depth of information (in terms of the choices available), links, and quality at the end of the task. For the many-items task (task 3), participants seem to have a higher percentage of mentions for layout and numbers at the start of the task, and higher mentions for content, pictures, and scope/depth at the end of the task.

Discussion

In this section we first present a summary of the major findings of this study, then discuss the implications of the results, and finally outline some limitations of this study.

Summary of Findings

Features and categories. By analyzing how searchers assess the utility of Web documents for information seeking tasks, we are able to construct a set of document features and feature categories that are employed for utility assessments. The five categories discovered (text, structure, quality, nontextual items, and physical properties) were discussed in detail in the section on Categories and Features. In contrast to previous work in analyzing user assessments of the utility of documents (e.g., Barry 1994, 1998; Cool et al., 1993; Schamber, 1991), our study focuses on Web documents rather than on formally structured, and generally high-quality, research articles. Moreover, in this work we did not predefine document features, as in Kelly et al. (2002); instead, we discovered the features from analyzing searchers' utility assessments.

Despite the difference in the structure and quality of Web documents and scientific articles, there is a large overlap between the features identified in our study and studies investigating research articles. Findings of a number of studies regarding the latter type of documents have been summarized in Maglaughlin and Sonnenwald (2002). Because of the particular nature of Web documents, a number of features were discovered that do not correspond to features of more formal document types. Such features mainly belonged to the physical properties category (e.g., file type, file size, connection speed, subscription/registration required). A significant contribution of our study stems from breaking down the text category into a number of contributing features (e.g., content, query terms, titles/headings, numbers) to gain a better understanding of the way searchers employ textual features in making relevance assessments. We did not gather data for other textual features (e.g., named entities, acronyms) that could automatically be detected and extracted by, for example, Web summarization or Web retrieval systems. A more complete analysis of textual features employed by searchers is something we intend to explore further in the future.

As far as the frequency of mentions is concerned, content was most frequently mentioned in our study (see Categories and Features and Table 2). More specifically, almost 45% of the total mentions of features corresponded to a feature from the text category. This is in agreement with findings of other authors (Maglaughlin & Sonnenwald, 2002).

However, in our study other aspects of documents were also frequently mentioned by searchers. Such aspects include structural elements of pages (e.g., layout of a page and hyperlinks in the page), quality aspects (e.g., scope/depth and the authority of the information contained in pages), as well as nontextual information contained in pages (i.e., pictures). Features corresponding to the structure and quality categories amount to almost 39% of the total mentions of features.

In the section Positive Versus Negative Mention of Features, we examined differences in mentions of features depending on whether they were used to indicate useful or nonuseful Web pages (positive and negative mentions, respectively). The results showed that certain features tend to indicate mainly useful pages (e.g., table data, links, authority/source, numbers), whereas others mainly nonuseful pages (Table 5) (e.g., links quality, general quality, page location, scope/depth of information). The analysis of the search sessions also revealed that documents that were assessed as useful often probed searchers to cite both positive and negative mentions of features (i.e., "this document is useful because of its layout and its content, but I don't like the fact that there are too many links in it"). This finding correlates to that of Maglaughlin and Sonnenwald (2002), who found that less than 50% of the documents that were judged relevant (or not relevant) in their study contained only positive (or only negative) mentions of features. This may have implications for the use of relevance feedback on the WWW, because feedback systems typically use the entire document as potentially relevant to the user's information need (White, Jose, & Ruthven, 2002). Our results suggest that it might be more beneficial if only highly useful parts of documents are considered by such systems.

Effect of task. In the section Variation of Feature Importance Across Search Tasks, we examined the variation of feature importance across the three information-seeking tasks used in the study. The data presented in Figure 2 and Table 6 demonstrate a considerable dependence of feature mentions on task characteristics. The results from the postsearch questionnaires also demonstrate that searchers' perceptions about aspects of the information seeking process varied for the various tasks. More specifically, the background task that requested participants to collect information about demographics of Internet users proved to be the most difficult of the three tasks. This task required users to synthesize information from different sources, making it harder for them to achieve a high perception of task completion. Participants had frequent mentions of content, links, numbers, and recency for this task.

The decision task (task 2) led participants to make frequent use of factual features of documents (numbers, pictures) and of the scope and depth of the available information. Participants also made less frequent use of links to access other pages of related interest. The many-items task (task 3) involved frequent mentions of links for the identification of "node pages" that would help users locate other pages with enough information to compile the required list of items. Participants for this task also made frequent use of query terms, pictures and the authority of the information. Participants for these two tasks had relatively similar perceptions of task completion as well as of other aspects of the information seeking process.

Progression of criteria along tasks. The data we presented in the section Progression of Criteria Along Tasks showed that participants' criteria along the duration of a task display a certain degree of variation. The degree of variation was larger for the second and third tasks, where participants had a higher perception of task completion. For the first task the variation was smaller, because the perceived difficulty of the task did not seem to allow participants to develop their information seeking strategies along the duration of the task.

Searchers' actions along the duration of a task suggest that they start by initially familiarizing themselves with the requirements of the task and the type of pages they had access to (e.g., those retrieved by a search engine). Then, during the process of the task, they identify those aspects of Web pages that would mostly help them complete the task. The type of features that become more frequently mentioned toward the end of a search session depend on the type of task. For example, for the decision task (task 2), searchers make more mentions of page layout and scope and depth of information toward the end of the task. Participants attributed certain importance to these two features after having examined other pages in relation to this type of task. This exposure led them to decide on what type of page layout they were looking for (i.e., one with pictures of the products, technical specifications, price details, and delivery options), as well as to opt for pages that provided them with a sufficient choice of products so that they could easily reach a decision by consulting as few pages as possible.

These results seem to be in general agreement with the findings of Vakkari (2000), who noted that in the initial stages of the search, participants seem to have a vague mental model of the task. This mental model gradually becomes more focused as they become familiar with the requirements of the task, and toward the end of the process participants seem to employ more focused and specific information. It should however be mentioned that Vakkari's study involved searches that span the duration of months rather than a single search session of 15 or 30 minutes as in our study. This difference may have affected the way participants evolved their search strategies.

Implications of Results

The results from this study have implications for the design of systems that use or recommend Web pages. We discuss such implications in this section.

The set of Web document features and feature categories that were discovered in our study, and the relative importance

attributed to them from the mentions of participants, provide an indication of which aspects of Web pages participants employ to make utility assessments of documents on the WWW. These features can be used to provide relevance clues to users of a search engine. Web document summarization systems (e.g., White et al., 2003) typically use only textual aspects of documents to inform searchers of the utility of retrieved Web documents. Other efforts to generate effective summaries of Web documents involve the creation of thumbnail previews of Web documents (e.g., Czerwinski, van Dantzich, Robertson, & Hoffmann, 1999; Dziadosz & Chandrasekar, 2002; Woodruff, Rosenholtz, Morrison, Faulring, & Pirolli, 2002) and the production of multimedia summaries of Web documents (Wynblatt & Benson, 1998). The increased effectiveness of these representations over textual summaries has been shown by Dziadosz and Chanrasekar (2002) and Woodruff et al. (2002). Thumbnail previews of Web documents give searchers an overview of the structure of a Web page, but they may not be of high enough quality to allow them to explore the content. Enhanced quality thumbnails (Woodruff et al., 2002) allow users to examine parts of the content of the pages, but they are more computationally expensive to create, an important aspect especially for the online generation of Web document summaries.

The set of features that were highly mentioned in this study may provide suggestions as to which document features should be considered by Web summarization systems. This may act as a less computationally expensive solution than thumbnail previews of documents, but a more effective one than systems that only use textual aspects of documents. It should be noted that some of the features ranked as important in this study can be easily automatically extracted by such a system (e.g., text, links, pictures), whereas some other features are more abstract in their definition and potentially more difficult to capture (e.g., scope/depth, overall quality). Ways to incorporate such aspects in Web document summaries need to be investigated. Vakkari (2000) has noted that by identifying what users expect from document representations, we can design more suitable representations. We view the research presented in this study as moving toward the direction suggested by Vakkari.

A further implication of this study can be in identifying nontextual aspects of Web pages that can be used in matching algorithms on the WWW (e.g., by search engines) for influencing the retrieval score of Web documents. Zhu and Gauch (2000), for example, used information quality metrics of Web pages to influence the retrieval of high-quality pages for Web searching. These quality factors included the authority of the pages (from reviews provided by ZDNET, 2003), the currency of the page (from the last date stamp), and the popularity of the page (how many other pages linked to the page). These metrics were used to investigate how aspects of page quality could improve retrieval effectiveness. Zhu and Gauch compared the effectiveness of matching algorithms based on page content against algorithms based on content plus quality metrics. They showed that incorporating noncontent features of pages could improve retrieval effectiveness over only considering content alone. However, the study by Zhu and Gauch did not examine how assessments of relevance can change according to the nature of information-seeking tasks, or how relevance assessments can change over the course of an individual search. Our results, by taking these two issues into account, can provide an indication of which features to use for different task types and for different stages within a search, to calculate a noncontent, quality score for pages that are to be retrieved.

The dependence of document features mentioned by participants on task type and on the stage of the search, suggest that there are also implications of this study for the interaction of searchers with Web documents. Web-based systems can take into account the type of search performed by an online user, and the stage at which the user is at his search, to weight differently the various features of Web pages. These weights can influence the type of Web pages recommended to searchers, or the type of Web page summaries (or thumbnails) that are presented to searchers. The successful identification of task type or stage of search is undoubtedly a big challenge. However, methods that take into account the structure of the user's query and how this changes during the course of a search may be effective at identifying task categories and task stages.

Limitations

A limitation of the methodology can be found in the subjective nature of transcribing and analyzing the participants' judgments. It may be the case that the experimenter's interpretation of a participant's judgment is not totally in line with the participant's intention. This is especially so with features that correspond to the text category (Table 2). It was in general hard to obtain a high level of detail about the textual features that users found as contributing to their assessment of Web pages. In some cases it was explicitly stated by users whether it was the titles, section headings, or other specific textual items of a Web page that they were using to make their judgments. However, in many cases it was not possible to further analyze what textual aspect users employed when making their assessments; such cases were recorded under the generic content feature. However, we do feel that the users' judgments were in general explicit and required little interpretation on the experimenter's behalf.

It should also be mentioned that the postsearch questionnaires may have prompted users to identify certain document features as useful versus not useful. For example, in the first task that each user performed maybe he was naïve regarding what kind of page features to mention. After being presented with the postsearch questionnaire, where users were asked to rate the importance of certain document features, users might have became more focused on identifying and rating these particular features. The rotation of the order in which tasks were performed may be seen as an attempt to counterbalance this effect.

A further issue stems from that the pages assessed by the participants were a direct consequence of the tasks, not a

representative set of Web pages. One implication of this is that some page features may be underrepresented in this sample, as was the case with multimedia features. A further implication is that the frequency of mentions of some features might have been influenced as much by the presence or not of these features in the returned document sets as by the importance of these features to searchers. Other researchers (Barry, 1994) have also noted that frequency of mentions can not automatically be equated with feature importance. However, one can also argue that the frequent presence of features in Web pages does in some way equate to the utility (as opposed to importance) of these features. For example, if text is present on most pages and searchers refer to it frequently, then it may be valid to say that text is a useful feature to identify. A more complete picture regarding feature presence and importance could have been obtained if there had been a recording of all features present in each Web page viewed by searchers in the study. For practical reasons this was not feasible, and consequently the results regarding the importance of features based on frequency of mentions should be seen in the context of the specific task types employed.

A further limitation can be found in that the search tasks used in this study were not real (i.e., did not reflect the participants' information needs). This may have affected the participants' behavior and criteria used when assessing the Web pages.

Conclusions

In this study, 24 participants were given three tasks each, and indicated the features of Web pages that they used when deciding about the usefulness of the pages in relation to the tasks. The tasks were presented to participants within the context of a simulated work-task situation. The major contribution of this work was threefold. First, it examined the features of Web pages that online searchers use when they assess the utility of Web pages for information-seeking tasks. Web page features were discovered as they were mentioned by the participants during the course of the tasks and were not previously defined by the authors. Second, it investigated how the features employed by the participants varied depending on the type of task. Third, it examined the variation of Web page features as participants progressed along the course of tasks. The findings of this study have implications for a number of issues on the design of systems that use or recommend Web pages.

Further work is necessary to examine a wider variety of task types and their characteristics, and the effect that they have on the features employed by users of Web pages. Because this study did not allow participants to search for a topic of their own interest, an issue to consider in the future would be the extension of this study in a more realistic setting where searchers pursue their own information needs. Searches could then be categorized to specific task categories, and feature mentions studied for each category. The incorporation of the features discovered in this study in an online WWW system, both for the retrieval of Web pages and for the presentation of summaries of Web pages to searchers, is also an issue that we plan to address in the future.

Acknowledgments

The authors wish to thank the members of the Glasgow IR group and of the University of Strathclyde e-communities group for helpful discussions and comments, as well as the reviewers for their helpful suggestions. The authors also wish to thank the searchers who participated in the study. This research is funded by the EPSRC (UK) research grant GR/R74642/01.

References

Barry, C.L. (1994). User-defined relevance criteria: An exploratory study. Journal of the American Society for Information Science, 45(3), 149–159.

Barry, C.L. (1998). Document representations and clues to document relevance. Journal of the American Society for Information Science, 49(14), 1293–1303.

Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation, 56(1), 71–90.

Chi, E.H., Pirolli, P., Chen, K., & Pitkow, J. (2001). Using information scent to model user information needs and actions on the Web. Proceedings of the ACM SIGCHI Conference (pp. 490–497).

Cool, C., Belkin, N.J., & Kantor, P.B. (1993). Characteristics of text affecting relevance judgments. In Proceedings of the 14th National Online Meeting (pp. 77–84). Medford, NJ: Learned Information.

Cothey, V. (2002). A longitudinal study of World Wide Web users' information-seeking behavior. Journal of the American Society for Information Science and Technology, 53(2), 67–78.

Czerwinski, M.P., van Dantzich, M., Robertson, G., & Hoffmann, H. (1999). The contribution of thumbnail image, mouse-over text and spatial location memory to Web page retrieval in 3D. In Proceedings of Interact '99 (pp. 163–170).

Dziadosz, S., & Chandrasekar, R. (2002). Do thumbnail previews help users make better relevance decisions about Web search results? In Proceedings of the ACM SIGIR Conference (pp. 365–366).

Fleming, J. (1998). Web navigation: Designing the user experience. Sebastopol, CA: O'Reilly & Associates.

Flick, U. (1998). An introduction to qualitative research. Thousand Oaks, CA: Sage.

Hosmer, D.W., & Lemeshow, S. (1989). Applied logistic regression. New York: Wiley.

Ingwersen, P. (1992). Information retrieval interaction. London: Taylor Graham.

Ivory, M.Y., Sinha, R.R., & Hearst, M.A. (2001). Empirically validated Web page metrics. Proceedings of the ACM CHI Conference on Human Factors in Computing Systems 3, 53–60.

Janes, J.W. (1991). Relevance judgments and the incremental presentation of document representations. Information Processing & Management, 27(6), 629–646.

Jansen, B.J., Spink, A., & Saracevic, T. (2000). Real life, real users and real information needs: A study and analysis of user queries on the Web. Information Processing & Management, 36(2), 207–227.

Kelly, D., Murdock, V., Yuan, X., Croft, W.B., & Belkin, N.J. (2002). Features of documents relevant to task and fact-oriented questions. Proceedings of the 11th ACM CIKM Conference (pp. 645–647).

Kim, K.S., & Allen, B. (2002). Cognitive and task influences on Web searching behavior. Journal of the American Society for Information Science and Technology, 53(2), 109–119.

Lazonder, A.W., Biemans, H.J.A., & Wopereis, I.G.J.H. (2000). Differences between novice and expert users in searching information on the World

Wide Web. Journal of the American Society for Information Science, 51(6), 576–581.
Maglaughlin, K.L., & Sonnenwald, D.H. (2002). User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. Journal of the American Society for Information Science and Technology, 53(5), 327–342.
Mizzaro, S. (1997). Relevance: The whole history. Journal of the American Society for Information Science, 48(9), 810–832.
Preece, J. (1994). Human computer interaction. Wokingham, UK: Addison Wesley.
Saracevic, T. (1969). Comparative effects of titles, abstracts and full text on relevance judgments. Journal of the American Society for Information Science, 22, 126–139.
Saracevic, T. (1970). The concept of "relevance" in information science: A historical review. In T. Saracevic (Ed.), Introduction to information science (pp. 111–151). New York: R.R. Bowker.
Schamber, L. (1991). Users' criteria for evaluation in a multimedia environment. Proceedings of the 54th Annual Meeting of the American Society for Information Science, 28, 126–133.
Schamber, L. (1994). Relevance and information behavior. Annual Review of Information Science and Technology, 29, 3–48.
Schamber, L., Eisenberg, M.B., & Nilan, M.S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. Information Processing & Management, 26, 755–776.
Search Engine Heaven. (2003). http://www.richeast.org/SEHeaven.html
Tang, R., & Solomon, P. (1998). Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior. Information Processing & Management, 34(2/3), 237–256.
Vakkari, P. (2000). Relevance and contributing information types of searched documents in task performance. In Proceedings of the 23rd ACM SIGIR Conference (pp. 2–9).
Vakkari, P., & Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. Journal of Documentation, 56(5), 540–562.
White, M.D., & Iivonen, M. (2001). Questions as a factor in Web search strategy. Information Processing & Management, 37, 721–740.
White, R.W., Jose, J.M., & Ruthven, I. (2002). The use of implicit evidence for relevance feedback in Web retrieval. In Proceedings of the 24th European Colloquium on IR Research (pp. 93–109).
White, R.W., Jose, J.M., & Ruthven, I. (2003). A task-oriented study on the influencing effects of query-biased summarisation in Web searching. Information Processing and Management, 39(5), 707–733.
Woodruff, A., Aoki, P.M., Brewer, E., Gauthier, P., & Rowe, L.A. (1996). An investigation of documents from the World Wide Web. Proceedings of the Fifth International World Wide Web Conference.
Woodruff, A., Rosenholtz, R., Morrison, J.B., Faulring, A., & Pirolli, P. (2002). A comparison of the use of text summaries, plain thumbnails, and enhanced thumbnails for Web search tasks. Journal of the American Society for Information Science and Technology, 53(20), 172–185.
WWW Digital Library. (2003). http://www.vlib.org/
Wynblatt, M., & Benson, D. (1998). Web page caricatures: Multimedia summaries for WWW documents. Proceedings of the IEEE International Conference on Multimedia Computing and Systems (pp. 194–199).
Yahoo. (2003). http://www.yahoo.com/
Yuan, W. (1997). End-user searching behavior in information retrieval: A longitudinal study. Journal of the American Society for Information Science, 48(3), 218–234.
ZDNET. (2003). http://www.zdnet.com/yil
Zhu, X., & Gauch, S. (2000). Incorporating quality metrics in centralized distributed information retrieval on the World Wide Web. Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 288–295).

Appendix

Instructions for Searchers

You will be given three tasks, similar to the one shown below.

> *"You are planning to attend a conference in Kyoto, Japan, in mid-September. The conference finishes on a Friday and you are planning to spend the weekend in Kyoto before flying back to the United Kingdom. To avoid wasting time whilst you are there, you want to get enough information to help you plan your free weekend before you go."*

You will be given a maximum of 30 minutes for each task. In order to tackle each task you will use the Internet. You are expected to use/browse the Internet as you normally would (i.e., use a search engine, navigate directly to a specific Internet address, etc.).

For each document that you choose to view as part of each task, we want to collect some information relating to how the document may help you complete the task at hand.

More specifically, the aim of this study is to see what criteria people use when deciding about the usefulness of Web pages. It is not your performance on the tasks that is measured, but rather what criteria you employ when deciding which web pages are useful in relation to a specific task.

To this end, we would like you to explain what features of each document that you view help you in completing the task (e.g., content, format/layout, source, multimedia content, etc.).

Your Internet session will be recorded automatically. This is to allow us to extract information relevant to this study.

During the process of the task you will be encouraged to explain the reasons for which you may have pursued a specific document, changed your query terms, etc. This will also help us gather useful information for the purposes of this study.

Presearch Questionnaire

1. Age:
2. Sex (Please circle)
   M / F
3. Occupation:
4. How often do you browse the Internet? Please circle the closest option.

Rarely	1–2 times a month	1–2 times a week	Once a day	Several times a day

5. How often do you use search engines to search the Internet? Please circle the closest option.

Rarely	1–2 times a month	1–2 times a week	Once a day	Several times a day

6. How often do you find what you are looking for when using search engines?

Very often				Not often at all
1	2	3	4	5

7. Which search engine do you mostly use?
   a. Altavista
   b. Google
   c. AllTheWeb
   d. Wisenut
   e. Other (Name):
8. Do you use advanced features of search engines (e.g., *More like this*, *phrase matching*, etc.)
   Y / N
9. Do you make use of bibliographic databases (e.g., Inspec, Medline, Citeseer etc.)?
   Y / N

Postsearch Questionnaire

1. How satisfied are you with the results of this search?

Very				Not at all
1	2	3	4	5

2. Do you think you have found enough information about the search task?
   Y / N
3. The search task we asked you to perform was:

	Very	Reasonably	Neither/Nor	Reasonably	Very	
Clear	1	2	3	4	5	Unclear
Complex	1	2	3	4	5	Easy
Familiar	1	2	3	4	5	Unfamiliar
Uninteresting	1	2	3	4	5	Interesting
Relevant to you	1	2	3	4	5	Not relevant to you

4. How stressful did you find the process?

Very				Not at all
1	2	3	4	5

5. How similar do you think your search behavior was during the previous task compared to your normal search behavior?

Very similar				Not at all similar
1	2	3	4	5

6. If you think your behavior was different, in what way did it differ to your normal search behavior?
7. For this task, how do you rate the importance of the following factors in helping you determine the usefulness of web pages to your search task:

	Very important				Not very important
Text	1	2	3	4	5
Multimedia	1	2	3	4	5
Links	1	2	3	4	5
Layout	1	2	3	4	5
Your previous knowledge of the topic of the search task	1	2	3	4	5
Who *produced* the page (the organisation/author)	1	2	3	4	5
The overall *quality* of the Web page	1	2	3	4	5
The *recency* of the page (how current is the information)	1	2	3	4	5
Other (name):	1	2	3	4	5

8. Have you got any other comments?

Final Questionnaire

1. Please place the tasks in order of how difficult they were to complete. (You may say more than one task was equally difficult/easy).

Most difficult	
Least difficult	

2. Do you have any general comments? (continue over the page if necessary)