Next Generation Web Search: Setting Our Sites

Marti A. Hearst
School of Information Management and Systems
University of California Berkeley
hearst@sims.berkeley.edu

Abstract

The current state of web search is most successful at directing users to appropriate web sites. Once at the site, the user has a choice of following hyperlinks or using site search, but the latter is notoriously problematic. One solution is to develop specialized search interfaces that explicitly support the types of tasks users perform using the information specific to the site. A new way to support task-based site search is to dynamically present appropriate metadata that organizes the search results and suggests what to look at next, as a personalized intermixing of search and hypertext.

1 Introduction

Surveys indicate that search engine user satisfaction has risen recently. According to one survey, 80% of search engine users say they find what they want all or most of the time [21]. This is a surprising result given average query length is still quite short – about 2 words. How can people be finding exactly what they want given such short queries?

Since no published large-scale analysis exists, we currently have to make guesses. I think the answer is that, as a gross generalization, most people use web search engines to find good starting points – home pages of web sites that discuss a topic of interest. A query on "horseradish" is very general; it does not indicate what it is the user wants to know about or do with horseradish, so the best thing a search engine can do is bring up sources of general information about horseradish which the user can then peruse in more detail.

This multi-stage process of search, starting with a general query and then getting more specific, is well-documented in non-web search [14]. Users of old search systems like Dialog and Lexis-Nexis were taught to first write a general query, look at how many (thousands) of results were returned, and then refine the query with additional terms until a reasonable number of documents resulted. In these older systems the user had to first select a collection, or source, to search. Many of these searchers were professionals, who knew a great deal about which sources were available after years of experience.

By contrast, in web search the purpose of the initial query seems primarily to be to choose the source – a web site of interest. Once at the source, the user has a choice of using hyperlinks to navigate through the many pages of information available, or using site search.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

In this article, I use the term "site" to mean a collection of information that has some kind of unified theme. This can be a collection of items such as architectural images, recipes, or biomedical texts, or the catalog of an e-commerce company, or the intranet of a company or a university (the theme being the various kinds of work done and information used by people in the organization), or what's begun to be called a "vortal" for vertical portal, which offers a wealth of different sources about one topic. FindLaw is an example of a vortal, providing search over dozens of different legal sources, including law journals, US Supreme court decisions, and legal news.

Major search engine companies are now offering site-specific search as part of their product suites. For example, Inktomi has announced plans to develop an architecture that allows flexible partitioning of information sets, allowing the assignment of weights to documents depending on which parts of the collection they occur in and what kinds of context they are associated with [19]. To best utilize this potentially powerful facility, an understanding is needed of how to combine this information effectively.

In the early days of web search, a query like "horseradish" would most likely have retrieved what felt like random pages. But two things have changed since then. First, more high-quality content has become available on a vast array of topics, and second, search engines now focus on returning definitive *sites*, rather than pages, for such general queries. For instance, the top hits for a search on "horseradish" on Lycos finds a link to the Horseradish Information Council, listed as part of the

Recreation > Food
Business > Industries > Food and Related Products > Fruit and Vegetables

web directory categories, and hits on encyclopedia entries for horseradish are shown alongside the

Reference > Encyclopedia > Microsoft Encarta > H
Reference > Encyclopedia > Encyclopedia.com > H

categories. These are followed by the home pages of various food products companies. (It is a pity that we cannot run this same query over the 1996 web using 1996 search engines for comparison purposes.)

The fact that search engines show hits on category labels is significant, because web directory categories, and their associated links, are manually selected to be representative starting points for search. The Google search engine is known for using hyperlink inlink information for ranking pages, based on the idea that if many pages link to a page, that linked-to page is likely to be of higher quality (recursively) because it is in effect "recommended" by the authors of the other pages. Others have also documented the merit of using inlink information for assessing page quality [12, 1], and other web search engines are making use of this information. However, Google also now incorporates category information into its search results, listing category labels beneath a search hit if that category had been assigned to it. An informal test on short general queries brought up on average 2.7 categories alongside the top 10 hits.[1]

A recent study [1] presents intriguing evidence that the *number of pages in the site* is a good predictor for the inlink-based ranking, at least for popular entertainment topics. In other words, the most popular sites on a topic according to inlink information are those sites that have a lot of information on the topic; the good sources. This bolsters the argument that what web search is good at doing is getting people to the right site or collection, after which the complicated information seeking begins.

2 Site Search and Tasks

In a well-designed web site, hyperlinks provide helpful hints about what is behind them and where to go next, a characteristic also known as "scent" [5, 16]. One usability expert claims that as a general rule, users do not object so much to following many links as they object to having to follow links that do not clearly support their task.

[1] The queries were horseradish, election, colon cancer, sex, berkeley, affirmative action, china, chat, recipes, united airlines.

If the user is unsure where to go next or has to resort to the back button, the usability of the site can decrease dramatically [20].

When the user has to resort to using the search facility, the results on a site are usually disorderly and shown out of context, and do not indicate what role the various hits play in the use or structure of the site. Usability gurus and ecommerce researchers alike lament the poor quality of site search, and claim that billions in business are lost each year due to poor site design and site search [20, 11].

What is a solution? Consider the following analogy. Following hyperlinks is like taking the train, whereas using site search is like driving an all-terrain four-wheel drive vehicle. On the train, there are a fixed number of choices of where to go and how to get there. To get from Topeka to Santa Fe you have to go through Frostbite Falls whether this make sense to you or not. On the other hand, you are unlikely to get lost – if you look at where you are on the map, all is clear. By contrast, a four-wheel drive Land Cruiser will take you anywhere, but you may get wedged between two boulders on the side of a cliff and be completely disoriented as to your whereabouts.

An ideal search interface adopts the best aspects of each technique. We would like a train system that magically lays down new track to suggest useful directions to go based on where we have been so far and what we are trying to do. The tracks follow the lay of the land, but cross over the crevasses and get to every useful part of the earth. They also allow us to back up at any time and take a different route at each choice point.

2.1 The Importance of the Task

Figuring out the best way to lay these tracks is nontrivial, because just as there is a huge variety of information available on these sites, there are also a huge variety of ways people use that information. To help cut down on the number of possible routes, the search interface, as well as the site structure, should reflect what is it people do with the site, that is, what *tasks* they attempt to accomplish on the site.

A recent study by Jared Spool's research firm uncovered the importance of task completion in user satisfaction in web site use [18]. The study compared 10 different sites; each participant conducted searches for information of interest to them on each site. Afterwards, Spool's team asked the participants to rate the web sites according to how fast they thought they were. Surprisingly, there was no correlation between the page download speeds and the perceived speed ratings. In fact, participants perceived the site with the fastest download speed to be the slowest, and vice versa.

However, there was a strong correlation between perceived speed and how successful the participants were at achieving their goals on the site. This was also correlated with how strongly the participant thought they usually "knew what to do next" at any given point in their task. Spool concludes that the correlational evidence suggests that if the goal is to improve perceived speed, it is more important to focus on designing web sites to help users complete their tasks, rather than focusing on task-neutral features like download speed.

I have now said that site search should reflect the tasks that a user would like to accomplish on the site, and I have suggested a metaphor about a magical train that lays tracks according to where you want to go and what you have done so far, a blending of the best features of hypertext and search. In the remainder of this section I will describe this idea in more detail and outline our research efforts in this direction.

2.2 Metadata

One more ingredient is needed before we can cook up this new idea. That is the notion of metadata. Metadata is commonly glossed as meaning "data about data". Most documents have some kinds of meta-information associated with them – that is, information that characterizes the external properties of the document, that help identify it and the circumstances surrounding its creation and use. These attributes include author(s), date of publication, length of document, publisher, and document genre.

Additionally, content-oriented subject or category metadata has become more prevalent in the last few years, and many people are interested in standards for describing content in various fields. Web directories such as

Yahoo and looksmart are familiar examples, and as seen above, web search engines have begun to interleave search hits on category labels with other search results.

Collections such as medical documents and architectural images have richer metadata available; some items have a dozen or more content attributes attached to them. It can be useful to think of category metadata as being composed of *facets*: orthogonal sets of categories, which together can be used to describe a topic. In the medical domain, the different facets are Disease type, Drug Type, Physiology, Surgery Type, Patient Type, and so on. Each article is a complex combination of several of these types of facets, each of which has a hierarchical structure. For example, a MedLine article entitled "Inhaled and systemic corticosteroid therapies: Do they contribute to inspiratory muscle weakness in asthma?" is assigned the MeSH categories *Steroidal Anti-Inflammatory Agents, Asthma, Muscular Diseases, Respiratory Muscles, Inhalation Administration, Adult, Beclomethasone, Case-Control Studies, Prednisone,* and *Risk Factors*, among others.

Researchers have long reported that using metadata in search is problematic, because the labels assigned often mismatch user expectations [15, 6]. Furthermore, category metadata is often inconsistent. Nevertheless, I believe the full potential of metadata in search results is underexplored and could yield significant improvements, especially for supporting task-based search over large collections of similar-style items (such as biomedical articles, architectural images, and recipes).

Metadata can be used as a counterpoint to free text, since free text and metadata can both be searched, but whereas free text queries must usually be subject to relevance ranking, metadata can be retrieved much as in a standard database query. It is much easier to accurately implement the query "Find all documents that have been assigned the category label Affirmative Action" than it is to implement "Find all documents about affirmative action".

2.3 An Example: epicurious

As a consequence of writing this paper, a website was brought to my attention that exhibits a good subset of the ideas I think can be useful for using metadata to improve task-oriented site search.[2] In this case the collection is recipe information; actually a database problem (fuzzy matching is not required) but many of the ideas can transfer to the fuzzier needs of information search.

Recipes are examples of information for which hierarchical faceted metadata is familiar to everyone. The facets used by epicurious are Main Ingredient, Cuisine, Preparation Method, Season/Occasion, and Course/Dish. Each of these has subcategories; for example, subcategories for Course/Dish include Appetizers, Bread, Desserts, Sandwiches, Sauces, Sides and Vegetables. The collection has over 11,000 recipes.

A standard search interface for recipes requires the user to either do a keyword search or drill down a category hierarchy. For example, on a different recipe site (called SOAR[3]), selecting Main Dishes > Poultry results in 57 recipes. To further refine these choices, some additional hyperlinked categories are shown:

> Poultry: Chicken Recipes; Poultry: Duck Recipes; Poultry: Game Hens;
> Poultry: Game Recipes; Poultry: Goose; Poultry: Turkey

Selecting Chicken retrieves a list of about 40 recipes and two more subcategories:

> Chicken Appetizers
> Diabetic Chicken Recipes

I could have also found the Chicken Appetizers category if I'd begun with the Appetizers category. However, if I'd wanted to see Italian recipes with chicken as a main dish, I have to first select Region > Italian, and then look

[2] http://www.epicurious.com/e_eating/e02_recipes/browse_main.html
[3] http://soar.berkeley.edu/recipes/

through all 665 Italian recipes. The site also allows a search for a particular term over the category, so I can do a search on "chicken" in the Italian section and hope for the best.

One problem with this interface is that I do not know how many recipes have a given attribute until after I follow the link. It would be much more useful to know this information when having to make a decision about where to go next. Researchers in the human-computer interaction community advocate the importance of information previews in this kind of situation [17]. Another flaw is the irregularity of what kinds of subcategories occur under a given category. Why are only appetizers shown as a special category, but not main dishes?

Recipe finding is in most cases a combination of a browsing and a search task, and so is a prime candidate for our ideas about combining the best of both techniques. The epicurious site does an excellent job of this. On this site, after selecting Main Ingredient > Poultry, the information of Figure 1 is shown.

Browse > Poultry

Refine by: Course/Meal | **Preparation** | **Cuisine** | **Season/Occasion**

Appetizers (65)	Brunch (5)	Main Dish (799)	Sauce (13)
Bread (1)	Condiments (2)	Salad (64)	Side (10)
Breakfast (1)	Hors d'Oeuvres (44)	Sandwiches (45)	Snacks (1)
Soup (73)	Vegetables (2)		

1 - 15 of 988 | **Next>**

1985 CHICKEN PIE WITH BISCUIT CRUST
Gourmet January 1991

ACAPULCO CHICKEN
Bon Appètit
...

Figure 1: A sketch of the epicurious site's method for browsing dynamic metadata describing a collection of recipes. After selecting a main dish type (poultry), the user can refine the results using four other types of metadata.

This view allows me to reduce the set of recipes along any of the other metadata facets. It also tells me how many recipes will result if I refine the current metadata facet (Course/Meal) according to one if the terms in its subhierarchy. If I select Hors d'Oeuvres, the view of Figure 2 results.

Note that I have created this combination of categories on the fly. I could have started with Course Dish > Hors d'Oeuvres, and then refined by Main Ingredient > Poultry and ended up at the same point.

Now I can further refine the set of recipes by selecting a preparation type, the subhierarchies for which are shown above. Alternatively, I can select Cuisine and the system will show the same set of 44 chicken Hors d'Oeuvres recipes according to cuisine type, such as Caribbean, Italian, Low Fat, and Kid-Friendly.

The site also allows search over document subsets. In the example above, I can search over the 44 chicken appetizer recipes to isolate those that include avocado or some other ingredient. Search is also allowed over the entire dataset. However, the results of search are shown as a long unordered list of recipes, and thus loses the benefits of the browsing interface. There is, however, an advanced search form that allows the user to select sets of main ingredients along with the other metadata types. It suffers in comparison to the browsing facility, however, in not helping users avoid empty or large results sets.

Browse > Poultry > Hors d'Oeuvres

Refine by: Preparation | **Cuisine** | **Season/Occasion**

Advance (4)	Broil (3)	Marinade (4)	Roast (3)
Bake (8)	Fry (2)	No Cook (1)	Saute (6)
Barbecue (4)	Grill (8)	Quick (6)	Slow Cook (1)

1 - 15 of 44 | **Next>**

BRANDIED CHICKEN LIVER PATE
Gourmet March 1996

BUFFALO WINGS
Epicurious January 1998
...

Figure 2: Result of revising the results of Figure 1 by selecting the Hors d'Oeuvres metadata type.

This interface supports information seeking for several different kinds of recipe-related tasks, including, e.g., "Help me find a summer pasta," (ingredient type with event type), "How can I use an avocado in a salad?" (ingredient type with dish type), and "How can I bake sea-bass" (preparation type and ingredient type). It does not support other tasks such as menu planning and organizing by customer reviews.

2.4 Example: Yahoo

The epicurious site provides a nice example of how to incorporate metadata into the search process, acting as a kind of dynamically-determined hyperlink. This is different from a setup like the directory structure at Yahoo, in which the paths are determined in advance. The Yahoo directory does combine certain types of metadata – most notably Region types are intermixed with other categories – but usually only up to two types are combined, and the combinations are not tailed to what the user wants to see. For example, to find UC Berkeley following links, the links that I clicked on were College and University > Colleges and Universities > United States > U > University of California > Campuses > Berkeley. This makes use of crosslinks (symbolic links), so the official category once I finally see a link for UC Berkeley is: U.S. States > California > Education > College and University > Public > University of California > Campuses. After clicking on Berkeley, the new category label that is actually associated with the information about UC Berkeley is U.S. States > California > Cities > Berkeley > Education > College and University > Public > UC Berkeley. This is an entirely different path than the one I traversed via hyperlinks. To handle the fact that most categories are best reached by multiple kinds of metadata, Yahoo does a great deal of crosslinking that causes the actual category labels traversed to change beneath the user. A system that lets the user choose which kinds of metadata to view next should be more effective.

By contrast to this clumsy use of metadata, Yahoo has a nice way of dynamically linking metadata in its restaurant selection site. At the top level it presents links that aid in tasks involving selection of restaurants, including maps and telephone directories. Before any searching can take place, the user must navigate a region metadata hierarchy to select a city. The user can then either select a cuisine link or issue a query over restaurant names or cuisine types, resulting in a list of restaurants that meet these criteria. After selecting a hyperlink for a particular restaurant, the user sees a summary of information about the restaurant which includes a set of links grouped under the label of "Find Nearby". These links include movies, bars and clubs, and cafes that can be found in the geographic neighborhood of the selected restaurant. (Researchers are also providing this type of functionality [4].) There is also a link labeled "More A&E" (A&E indicates Arts and Entertainment, the supercategory for Restaurants and for Movies), but this breaks the conceptual model by showing a listing of all entertainment choices in the city, not those limited just to be near the selected restaurant.

In essence, Yahoo has assumed that many of those users searching the restaurant collection are actually engaged in a larger task which can be paraphrased as the stereotypical "find evening entertainment" task: dinner and a nearby movie. Two metadata facets are combined here: the region facet and the entertainment facet, and within this, two subhierarchies with the entertainment facet have been linked together – restaurant (with a cuisine attribute) and movies.

2.5 Integrating Search

The epicurious site does a nice job of dynamically suggesting metadata to help the user reduce the set of documents in an organized manner, making use of information previews to show how many documents would result after each choice, and allowing the user to easily back up to earlier states in the search process by clicking on the hyperlinks indicating the path taken so far. However, the interface does not interweave the search results into the category structure, and a straightforward improvement would be to organize the search results according to the same category metadata layout that is used for the browsing interface.

However, a large text collection such as the MedLine collection of biomedical abstracts is more difficult to search and organize than something like recipes. Additional facilities may be needed to apply this kind of dynamic metadata interface to something as complex and voluminous as medical text, and information retrieval-style ranking is probably necessary along with keyword search to help sort through results. Furthermore, the metadata is in some cases more hierarchical than in the recipe example, and more types of metadata are available, so only a subset should be shown at any given time. The system should dynamically determine which *types* of metadata to show, based on what the user has done so far and their past history (this idea has been pursued in other contexts [10, 13]). For example, a clinician prescribing medications for a patient may want to be able to always see categories associated with this patient's particular allergies.

2.6 Example: BioMedical Text

Consider the following example. Say a medical clinician named Dr. Care needs to find information about the use of cortisone shots as a treatment for asthma for a particular patient. Using our proposed system, Dr. Care can begin either by selecting an initial category label or by typing in some terms directly. Assume she already knows the MeSH category labels for the high-level concepts *Asthma* and *Steroids* but does not want to have to remember the category labels for more specific terms. By specifying these labels directly, the equivalent of a conjunctive Boolean search is run over the collection.[4] This returns 99 articles, which have a total of 2000 MeSH subject headings, 577 of which are unique.

Figure 3 shows an example of what an interface that incorporates the ideas discussed above might look like. The system provides Dr. Care with a way to get started in dealing with this large result set. It indicates the path taken to get to this point, the titles of two of the articles, suggestions for next steps below this, and the full document list at the bottom.

At the top is shown a hyperlinked path indicating the two choices made so far. The links allow the user to easily go back to an earlier stage in the navigation process; by selecting the Asthma link, Dr. Care would see the

[4] The data for this example was generated by running queries over the Medline/Healthstar database for 1995-1999 as of July, 1999, using the California Digital Library interface to this system. The collection contains article abstracts from 8,400 journals. Article information was downloaded and processed by hand in order to obtain the numbers. In some cases the details are simplified for expository purposes.

Asthma > Steroids

1. *A steroid-induced acute psychosis in a child with asthma* [Review]
2. *Management of steroid-dependent asthma with methotrexate* [Meta-analysis]

Steroids	**Other Views**	**User-Preferred**
▹ Pregnanes	◇ Admin. & Dosage (50)	◇ Musculoskeletal (4)
▹ Pregnadienes (5)	◇ Drug Effects (20)	◇ Drug Resistance (6)
▹ Prednisone (5)	◇ Therapeutic Use (25)	
▹ Pregnenes	◇ Risk Factors (4)	
▹ Budesonide (4)	◇ More ...	
▹ Corticosterone (3)		◇ All Categories (99)

99 Documents: [Sort by author] [Sort by popularity] [Sort by Steroids] [Cluster]

1. Effect of short-course budesonide on the bone turnover of asthmatic children
2. Effect of prednisone on response to influenza virus vaccine in asthmatic children.
...

Figure 3: Sketch of a proposed method for browsing and searching biomedical text.

documents associated with this term, independent of Steroids.

Below this appears a box showing a few selected document titles. These are a review article and a meta-analysis, which are appropriate given the high-level terms used so far in the navigation process. Other reviews exist in this result set but talk about specific steroids and so are not shown at this point. This illustrates another important idea: a search interface should match the level of generality of the retrieved documents to the generality of the current navigation state. For example, at the early stages of the search, overview articles should be shown, but as the search becomes more specific, so should the documents.

Below the title box are shown lists of categories. The lefthand side shows a subset of the Steroids portion of the chemicals facet of MeSH. Only those categories within the Steroids subtree that occur significantly within the 99 documents of this result set are shown. The steroid metadata hierarchy implicitly shows which branch of the steroid family they occupy, along with the number of documents in this subcollection that refer to the particular steroid. Thus the user sees a preview of what would happen if they added any of these terms to the query; the results would be reduced dramatically. But rather than actually issuing a query, the user can simply click on the category label, thereby navigating to a portion of the search space that contains the steroid term conjoined with the other terms used so far. The user can subsequently easily back up to the current state by following the hyperlinked path at the top of the screen.

The righthand column shows how user-preferred categories can be integrated into the interface. Say Dr. Care is concerned about musculoskeletal and drug resistance issues for this particular patient. These categories will appear on all views of the results collection, along with a preview of how many documents are in the set. These preferences could be specified directly by the user, inferred from previous selections, or based on citation structure or popularity in terms of page accesses. Finally, the user can elect to see all categories that correspond to documents in the current set, if they prefer to override the system's organization facilities.

The bottommost portion of the display allows Dr. Care to scan the retrieval results directly, and also allows her to reorder the titles in various ways. Sorting by popularity takes into account how often other users have viewed the documents. It also provides a "Sort by Steroids" option. This appears because Steroids is the most recently chosen category (which is also why it appears in the lefthand category column). The resulting ordering would make use of the structure of the Steroids subhierarchy to group documents with similar steroids together. Finally, users can invoke a clustering option to show a textual or graphical display of documents according to their overall commonality [9].

As these examples from disparate fields show, there are commonalities in supporting effective search strategies between domains. In our new research project, FLAMENCO, we are investigating these ideas of FLexible information Access using MEtadata in Novel COmbinations. Our end goal is to develop a general methodology for specifying task-oriented search interfaces across a wide variety of domains and tasks. We suggest that rich, faceted metadata be used in a flexible manner to give users information about where to go next, and to have these suggestions and hints reflect the users' individual tasks.

3 Other Approaches to Site Search

3.1 Specialized Interfaces

Another way to improve web site search is to create a specialized interface that takes the structure of the information on the site into account. One of our research projects applies a variation of this idea in an attempt to improve Intranet search. Intranets contain the information associated with the internal workings of an organization, and our system, called Cha-Cha, organizes web search results in such a way as to reflect the underlying structure of the organization. An "outline" or "table of contents" is created by first recording the shortest paths in hyperlinks from root pages to every page within the web intranet. After the user issues a query, these shortest paths are dynamically combined to form a hierarchical outline of the context in which the search results occur [3]. For example, hits on the query "earthquake" will be shown to fall within the mechanical engineering department, a science education project, and administrative pages that indicate emergency evacuation plans.

This interface has been deployed as the UC Berkeley site search engine for the last two years, receiving on average about 200,000 queries a month. Based on user interviews and surveys, it is sometimes quite useful to see the context in which the search hit occurred, especially when the query does not produce a good hit. On the other hand, the extra information can be overwhelming and unnecessary if the search is relatively straightforward. Furthermore, an Intranet's structure does not always reflect the user's task; often some other kind of organization would be more appropriate. For example, a user trying to find out about research on the effects of old-growth logging probably cares about the different kinds of logging under consideration, but not which university department the results came from.

Another example of a search interface that follows the structure of the information is the CiteSeer interface [7] which focuses on a *type* of information – scientific references – rather than a content domain. The search structure reflects the structure of the underlying information: citations have hyperlinks to other papers by the same authors, and to paragraphs of text of articles in which the target article is cited. Search results of aggregates of users are exploited both for ranking and for producing informative statistics, such as how many articles cite a particular author's papers. The interface shows special links that would not make sense in general search, and are tailored to what people searching research literature are interested in.

3.2 Question Answering

I believe a number of other approaches to web search will flourish as time goes on. These include rent-an-expert sites, where users are matched up with people who have expertise in a field and answer their questions for a fee.

Some services connect clients and experts via the phone, avoiding the necessity of typing and having the potential to spread easily to use via mobile devices (Keen.com and Exp.com are two examples).

Systems to automate question answering are also improving, and these will become important supplements for organizations' web sites, to handle customer questions more quickly and cheaply. Rather than generating an answer from scratch, these systems attempt to link a natural language query to the most pertinent sentence, paragraph, or page of information that has already been written. They differ from standard search engines in that they make use of the structure of the question and of the text from which the answers are drawn. For example, a question about what to do when a disk is full needs to be linked to an answer about compressing files or buying new disk. A user asking "Why does my computer keep hanging?" wants to find information about how to avoid this situation, whereas a user asking "How do I get my computer to print?" wants information about how to bring the situation about. The syntax of the question, as well as the content words, determines the kind of acceptable answer.

Question answering in a limited domain can be very powerful but it is much harder in a broad domain. Sophisticated question answering has not yet gotten far in web search aside from the well-known example of AskJeeves, in which question types are manually linked in advance to specific answer pages. However, researchers [2, 8] and companies (such as AnswerLogic and InQuizit) are developing domain-specific natural language processing algorithms and lexical resources that should greatly improve automated question answering in the next two to three years.

Technologists are becoming highly interested in what might be considered "real-world" metadata. The thinking is that a query of "Where is a good Mazda mechanic?" should automatically take note of the local time and location of the question asker, in order to make suggestions of car repair places that are both nearby and open at the time the question is asked. The need for such context-aware questions answering systems can be expected to grow along with the demand for networked mobile devices.

4 Integration with General Web Search

Returning to the original topic of this essay, what is the role of general search engines in the framework proposed above? General search engines should evolve to direct people to task-oriented solutions, instead of collection-oriented solutions as the modus operandi today. In other words, search engines will need to match user requests to task descriptions. One step in this direction is to interpret multi-word queries in terms of their implicit task. For example, to use document genre to determine which results to return. As a straightforward example, a query on "review" alongside another term such as the name of a play should bring back sites with theatre reviews. Currently the unstated default for most queries, for ad servers at least, is that the user's task is to buy something. Eventually I envision task directories supplementing directories, and search engines providing search over these descriptions.

Acknowledgements
I would like to thank Jan Pedersen for many helpful conversations about web search, Doug Cook for detailed information about Inktomi's current search results statistics, and Eric Brewer, Ame Elliott, Dan Glaser, Luis Gravano, Jason Hong, Rashmi Sinha, and Hal Varian for helpful comments on this article.
This research was supported an NSF CAREER grant, NSF9984741.

References

[1] Brian Amento, Loren Terveen, and Will Hill. Does 'authority' mean quality? Predicting expert quality ratings on web documents. In *Proceedings of the 23rd Annual International ACM/SIGIR Conference*, pages 296–303, Athens, Greece, 2000.

[2] Claire Cardie, Vincent Ng, David Pierce, and Chris Buckley. Examining the role of statistical and linguistic knowledge sources in a general-knowledge question-answering. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000)*, pages 180–187. Association for Computational Linguistics/Morgan Kaufmann, May 2000.

[3] Michael Chen, Marti A. Hearst, Jason Hong, and James Lin. Cha-cha: A system for organizing intranet search results. In *Proceedings of the 2nd USENIX Symposium on Internet Technologies and Systems*, Boulder, CO, October 11-14 1999.

[4] Junyan Ding, Luis Gravano, and Narayanan Shivakumar. Computing geographical scopes of web resources. In *Proceedings of the Twenty-sixth International Conference on Very Large Databases (VLDB'00)*, Sept 2000.

[5] George W. Furnas. Effective view navigation. In *Proceedings of ACM CHI 97 Conference on Human Factors in Computing Systems*, volume 1 of *PAPERS: Information Structures*, pages 367–374, 1997.

[6] Fredric Gey, Hui-Min Chen, Barbara Norgard, Michael Buckland, Youngin Kim, Aitao Chen, Byron Lam, Jacek Purat, and Ray Larson. Advanced search technologies for unfamiliar metadata. In *Meta-Data '99 Third IEEE Meta-Data Conference*, Bethesda, MD, April 1999.

[7] C. Lee Giles, Kurt Bollacker, and Steve Lawrence. CiteSeer: An automatic citation indexing system. In *Digital Libraries 98 - The Third ACM Conference on Digital Libraries*, pages 89–98, Pittsburgh, PA, June 1998.

[8] Sanda Harabagiu, Marius Pasca, and Steven Maiorano. Experiments with open-domain textual question answering. In *Proceedings of the COLING-2000*. Association for Computational Linguistics/Morgan Kaufmann, Aug 2000.

[9] Marti A. Hearst, David Karger, and Jan O. Pedersen. Scatter/gather as a tool for the navigation of retrieval results. In Robin Burke, editor, *Working Notes of the AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*, Cambridge, MA, November 1995. AAAI.

[10] Haym Hirsh, Chumki Basu, and Brian D. Davison. Learning to personalize. *Communications of the ACM*, 43(8), Aug 2000.

[11] Mark Hurst. Holiday '99 e-commerce. http://www.creativegood.com, Sept 1999.

[12] Jon Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[13] Henry Lieberman. Letizia: an agent that assists web browsing. In *Proceedings of 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 924–929, 1995.

[14] Gary Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, 1995.

[15] Karen Markey, Pauline Atherton, and Claudia Newton. An analysis of controlled vocabulary and free text search statements in online searches. *Online Review*, 4:225–236, 1982.

[16] Peter Pirolli. Computational models of information scent-following in a very large browsable text collection. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 3–10, Vancouver, Canada, May 1997. ACM.

[17] Catherine Plaisant, Ben Shneiderman, Khoa Doan, and Tom Bruns. Interface and data architecture for query preview in networked information systems. *ACM Transactions on Information Systems*, 17(3):320–341, 1999.

[18] Tara Scanlon, Will Schroeder, Richard Danca, Nina Gilmore, Matthew Klee, Lori Landesman, Amy Maurer, Paul Sawyer, and Jared Spool. *Designing Information-Rich Web Sites*. User Interface Engineering, 1999.

[19] Chris Sherman. Inktomi inside. http://websearch.about.com/internet/websearch/library/weekly/aa041900a.htm, April 2000.

[20] Jared Spool. *Web Site Usability: A Designer's Guide*. Morgan Kaufmann, 1998.

[21] Danny Sullivan. NPD search and portal site study. http://searchenginewatch.internet.com/reports/npd.html, July 6 2000. NPD's URL is http://www.npd.com.