Appearing in *The Sixth International World Wide Web Conference*, April 1997.

ParaSite: Mining Structural Information on the Web

Ellen Spertus
MIT Artificial Intelligence Lab and University of Washington Dept. of CSE
University of Washington
Box 352350
Seattle, WA 98195-2350
ellens@ai.mit.edu

Abstract

Web information retrieval tools typically make use of only the text on pages, ignoring valuable information implicitly contained in links. At the other extreme, viewing the Web as a traditional hypertext system would also be mistake, because heterogeneity, cross-domain links, and the dynamic nature of the Web mean that many assumptions of typical hypertext systems do not apply. The novelty of the Web leads to new problems in information access, and it is necessary to make use of the new kinds of information available, such as multiple independent categorization, naming, and indexing of pages. This paper discusses the varieties of link information (not just hyperlinks) on the Web, how the Web differs from conventional hypertext, and how the links can be exploited to build useful applications. Specific applications presented as part of the ParaSite system find individuals' homepages, new locations of moved pages, and unindexed information.

Introduction

The World-Wide Web contains millions of pages of data. Practical access to this information requires applying and expanding hypertext research to build powerful search tools. Most Web search tools only make use of the text on a page, ignoring another rich source of information, the links among pages. Much human thought has gone into creating each hyperlink and labeling it with anchor text. Other valuable relational information can be gleaned from the structure, hierarchy, and similarity of pieces of text. This information is already used by individuals when they browse the Web. It should be harnessed to build powerful automatic search tools.

Hypertext research has primarily focused on a single document or set of related documents converted to hypertext by a single individual, team, or program. We will refer to such a document or collection as "classical" hypertext. While the Web bears some similarity to a classical hypertext collection -- each is a collection of pages connected by hypertext links -- there are important differences, requiring a broader view of hypertext. The remainder of this section discusses ways in which the Web differs from classical hypertext, outlining new problems and opportunities. The next section discusses the types of links on the Web (not just hyperlinks) and heuristics for exploiting them, including extending the notion of collaborative filtering [Shardanand and Maes 1995]. The following section discusses how the heuristics could be applied to build tools to perform useful tasks, such as finding individuals' homepages, new locations of moved pages, and unindexed information.

Differences between the Web and Classical Hypertext

Links across documents and sites: While classical hypertext systems have links, they are qualitatively different from those on the Web, since they don't cross site boundaries and often don't even cross document boundaries. Classical hypertext documents or collections are limited to a single topic (or set of related topics). In contrast, techniques are needed for inferring the topics of Web pages. This problem is exacerbated not only by the lack of standards in labeling pages and links but also due to a new motive for deception. Opportunists now place misleading information on their Web pages in order to garner more visitors, such as by including spurious keywords to trick a search service into listing a page as rating highly in a popular subject. In old-fashioned hypertext, built by a single individual or team, such deception was nonexistent. We will show that these differences can be dealt with by relying less on how a page is labeled by its authors and by taking advantage of independent labeling.

Repeated or missing information: Another difference is that the Web is simultaneously redundant and incomplete. A hypertext manual probably answers each question in exactly one place or in none, making it crucial that no likely solution be overlooked. In contrast, an answer could appear any number of times on the Web, making recall less crucial. Consequently, ignoring pages that are difficult to analyze is more feasible with the Web than with classical hypertext. One disadvantage of the Web, however, is that the absence of a hyperlink between two pages does not imply they are unrelated, as it would in a traditional hypertext document [Frei 1992].

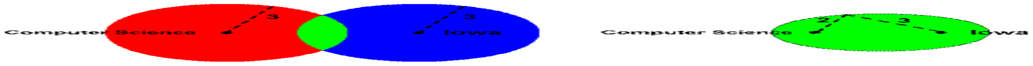
Constant change: Unlike classical hypertext, the Web is constantly changing, creating the new problem of finding information that has not yet been indexed. Furthermore, information that once was available might have been moved or deleted, leaving the user with a broken URL (Universal Resource Locator). On the positive side, the change on the Web gives it an extra dimension -- time -- that can be exploited by accessing old data stored by search engines and comparing it to up-to-date versions of pages. The old data is no more obsolete than a history book; both contain information about the past that may be useful. As we will show, while URLs are not as robust as URNs (Universal Resource Names) [Andrews 1996], we can exploit the additional machine, directory, and file name information in URLs.

Mining Links

While we have been using "links" to refer to hyperlinks, there exist other types of links, such as those in a directory hierarchy, in the structure of a document, or in a bibliographic citation index. Making use of the neglected information in all these kinds of links will allow the construction of more powerful tools for answering user queries.

Naïve Link Geometry

Consider the set of pages within three links of a computer science index. The pages within one link of the index are almost certainly related to computer science; the pages two or three links out have a lower probability, although still greater than that of random pages on the web. We can think of the set of pages within *n* links of an index *I* as being the interior of a circle (hypersphere) of radius *n* with center *I*. We could create two such circles with centers (foci) representing different subject indices and intersect them in an attempt to find material that is relevant to both subjects, as shown in Figure 1a:



(a) overlapping circles

(b) an ellipse

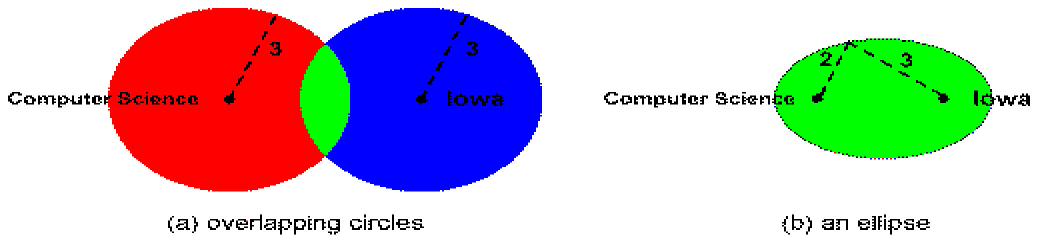


Figure 1: Finding Pages Relevant to Two Topics

(Intersection is probably not the ideal operation, since it could exclude a page very close to one focus but a little too far from the other, while including a page the maximum distance from each focus. Another option is to take the set of points where the sum of their distances from the foci is less than a constant; in other words, an ellipse, as shown in Figure 1b.)

Indeed, crawls from Yahoo's Computer Science (CS) and Iowa indices meet, appropriately, at Grinnell (Iowa) College's Department of Mathematics and Computer Science (www.math.grin.edu), showing that link geometry can be a useful technique for finding pages on a given set of topics. Unfortunately, not all forays are this successful. Popular pages with no relation to the desired topics, such as Lycos, are frequently returned, as are pages only tangentially related. For example, crawls from the MIT Artificial Intelligence Lab homepage and from Yahoo's CS index met at the Unified Computer Science Technical Report index (www.cs.indiana.edu/cstr/search).

Hypertext Links

Heuristics

The variation in quality of the pages reached through naïve link geometry suggests that not all links are equally useful. While many people have advocated explicitly typing links [Shum 1996], typed links are not in widespread use. In their absence, some inference is possible from syntactic information appearing in links [Pirolli et al. 1996]. This relies on file hierarchy information usually being available from the URL (i.e., the directory and file name). As Figure 2 shows, hypertext links within a site can be *upward* in the file hierarchy, *downward*, or *crosswise*. We refer to links to other sites as *outward*. In the crawls described above, we did not follow upward links in Yahoo, which would have led to more general topics.

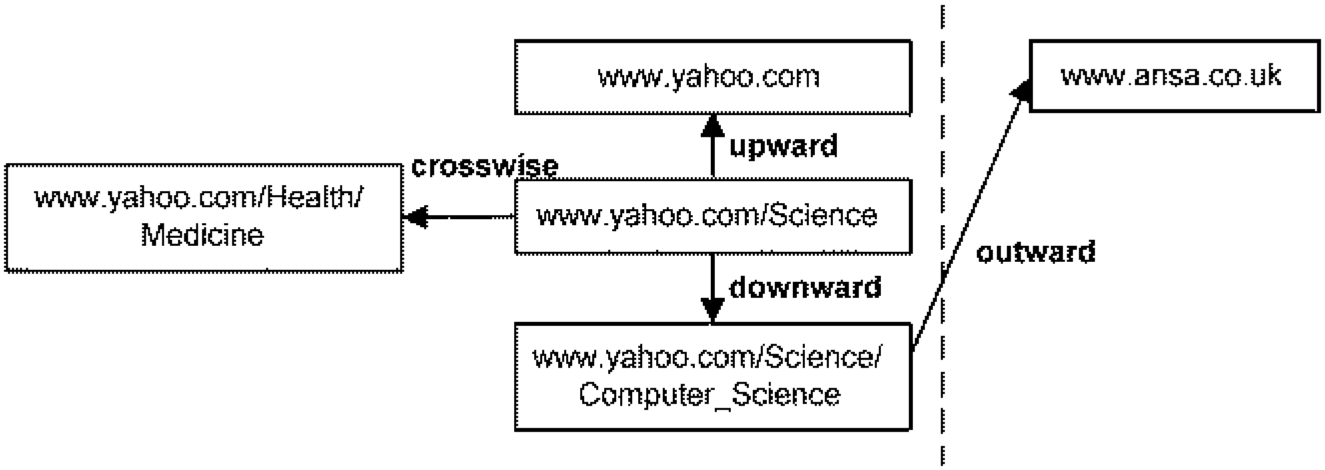


Figure 2: Directions of hypertext links. Links on the same server can be upward or downward in the file hierarchy or crosswise. Links to different servers are considered outward.

We can use these link categories to guess the type of a page. For example, a topic index typically contains many outward links, densely placed. Institutions' homepages usually also contain many links, but most of them downward. We can infer other information about a page from the ratio of links to it and from it. For example, we might guess a page to be popular if it has more links toward it than from it (although index pages would have trouble meeting this criterion). Note that pages have both topics (such as computer science) and types (such as homepage, index, or Yahoo page).

The most common types of links on Yahoo pages are downward links to subcategories and outward links to instances. Any page reached by following only downward links ($\overset{D}{\rightsquigarrow}$ *) indicates a specialization of the original page's topic; e.g.,*

> *Yahoo:Science* $\overset{D}{\rightsquigarrow}$ *Yahoo:Science:Computer Science* $\overset{D}{\rightsquigarrow}$ *Yahoo:Science:Computer Science:Artificial Intelligence*

In this case, the subcategorization relationship is also evident from the titles and URLs of the pages. This is not the case with crosslinks ($\overset{C}{\rightsquigarrow}$*) within Yahoo, which also indicate specialization. For example, one can traverse the following sequence of hypertext crosslinks:*

> *Yahoo:Science* $\overset{C}{\rightsquigarrow}$ *Yahoo:Arts:Humanities:History:Science and Technology* $\overset{C}{\rightsquigarrow}$ *Yahoo:Computers and Internet:History*

This leads us to the following heuristic:

> ***Heuristic 1 (Taxonomy Closure):*** *Starting at a page within a hierarchical index, following downward or crosswise links leads to another page in the hierarchical index whose topic is a specialization of the original page's topic.*

We also can guess the topic of a page reached through an outward link from an index:

> ***Heuristic 2 (Index Link):*** *Starting at an index, any page reached by following a single outward link is likely to be on the same topic.*

In general, we cannot know the topic of a page reached through multiple outward links. An exception is if the first outward link takes us to an index, in which case we can repeatedly apply Heuristic 2:

> ***Heuristic 3 (Repeated Index Link):*** *Starting at an index P and following an outward link to index P', a page reached through a further outward link is likely to be on the same topic (or a specialization of it) as the original page P.*

We can use these heuristics to evaluate the pages returned from a naïve crawl.

A predictably succesful crawl: *Let us first examine the path from Yahoo's Iowa index to Grinnell's Department of Math and Computer Science:*

> Yahoo:Regional:U.S. States:Iowa $\overset{O}{\rightsquigarrow}$ Iowa Internet Sites $\overset{O}{\rightsquigarrow}$ Grinnell's Department of Math and Computer Science

Because "Iowa Internet Sites" is an index, we can apply Heuristic 3 (Repeated Index Link) to conclude that the topic of Grinnell's Department of Math and Computer Science is a subtype of Iowa. Now, let us look at the path from Yahoo's computer science index:

> Yahoo:Science:Computer Science $\overset{D}{\rightsquigarrow}$ Yahoo:Science:Computer Science:Institutes $\overset{O}{\rightsquigarrow}$ Grinnell's Department of Math and Computer Science

By applying Heuristic 1 (Taxonomy Closure) and then Heuristic 2 (Index Link), we guess that the topic of Grinnell's Department of Math and Computer Science is a subtopic of computer science. Since its topic is (or can be seen as) "computer science in Iowa," both of our conclusions are correct.

A predictably unsuccessful crawl: Let us now look at a less useful intersection. The first common descendant of the University of Washington Computer Science and Engineering page and Yahoo's Iowa page was Lycos, through these paths:

> University of Washington Computer Science and Engineering page $\overset{D}{\rightsquigarrow}$
> Desktop References $\overset{O}{\rightsquigarrow}$ Lycos
> Yahoo:Regional/U_S__States/Iowa/ $\overset{O}{\rightsquigarrow}$ Iowa PROfiles $\overset{O}{\rightsquigarrow}$ Lycos

Because "Desktop References" and "Iowa PROfiles" are not topic indices, we can infer nothing about the page they point to.

Note that the topic of a page can be inferred not just from the path by which it was found but by discovering other ways in which it can be reached. For example, whether or not a page was originally found through Yahoo, one could guess its topic by tracing "backlinks" from it (i.e., pages that point to it) until Yahoo is reached.

Directory Links

The categorization of hyperlinks into upward, downward, crosswise, and outward makes use of information in the URLs about the relative locations in the file hierarchy of linked pages. It is also possible to consider the directory structure relation of pages in the absence of hypertext links. For example, in the file hierarchy, the directory "www.ai.mit.edu/people/ellens/" is a parent of "www.ai.mit.edu/people/ellens/Family". Whether or not there are hypertext links between the files in the two directories, there are links in the directory tree. We might make the following inference:

> **Heuristic 5 (Authorship Location):** If P is a homepage and file P' is in a directory below that of P, then P' is likely to be authored by the person identified on page P.

In general, there is a correlation between hypertext links and directory links. For example, a person's home page usually points to documents he or she created, many of which are stored in the person's directory (or its subdirectories). These pages typically point back to the home page, sometimes indirectly. We can state this *as:*

> ***Heuristic 6 (Directory/Hyperlink Correlation):*** *If page P is above page $P^{/}$ in the file hierarchy, one can probably follow hyperlinks from P to $P^{/}$ and vice versa. This is especially likely if P is a home page.*

Another way that information in URLs can be exploited is by looking for stereotypical directory names. For example, URLs for homepages often have the penultimate directory named "homes", "users", or "people". The final directory is often the same as the user's account name, and the file name often has as its base "index", "home", "default", or the user's account name.

Structure within a Page

While the Web could be represented as a graph with nodes representing pages and arcs representing hyperlinks, doing so would ignore much valuable information. Web pages should not be viewed as atomic objects. Each page is a loosely-structured collection of text and links, as shown in Figure 3:

Ellen Spertus's Personal Page

Return to my home page

Family

Relatives with Web pages:

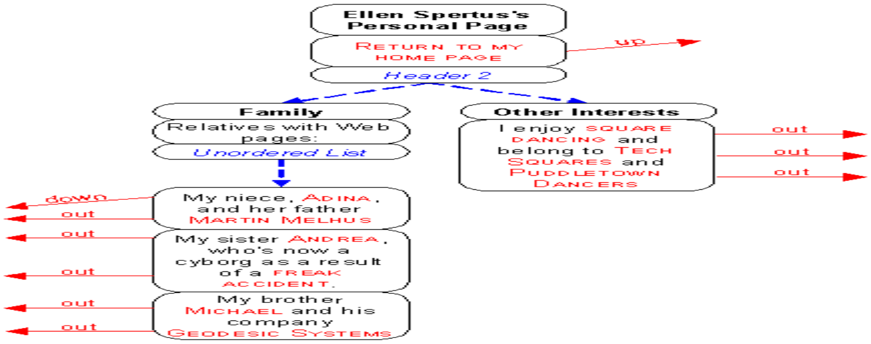
- *My niece, Adina, and her father Martin Melhus*
- *My sister Andrea, who's now a cyborg as a result of a freak accident.*
- *My brother Michael and his company Geodesic Systems.*

Other Interests

I enjoy square dancing and belong to Tech Squares and Puddletown Dancers.

Figure 3: A personal home page

Instead of treating the page shown in Figure 3 as a single node, it can be considered a tree of nodes, each with attached text and links embedded in the text, as shown in Figure 4:


Ellen Spertus's Personal Page
Return to my home page
up
Header 2
Family
Relatives with Web pages:
Unordered List
Other Interests
I enjoy square dancing and belong to Tech Squares and Puddletown Dancers
out
out
out
down
out
My niece, Adina, and her father Martin Melhus
out
out
My sister Andrea, who's now a cyborg as a result of a freak accident.
out
out
My brother Michael and his company Geodesic Systems

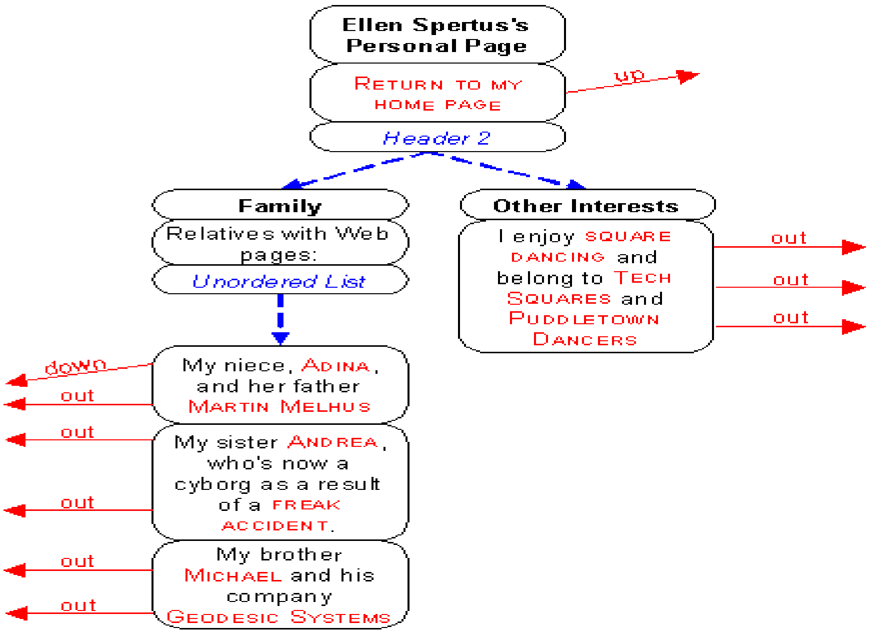

Ellen Spertus's Personal Page
Return to my home page
up
Header 2
Family
Relatives with Web pages:
Unordered List
Other Interests
I enjoy square dancing and belong to Tech Squares and Puddletown Dancers
out
out
out
down
out
My niece, Adina, and her father Martin Melhus
out
My sister Andrea, who's now a cyborg as a result of a freak accident.
out
out
My brother Michael and his company Geodesic Systems
out

Figure 4: The structure of the page shown in Figure 3

Links within the same list are more closely related than those in separate lists, and links within the same item (such as to Adina and to her father) will tend to be most closely related. In other words,

> ***Heuristic 7 (Spatial Locality of References):*** *If URLs U_1 and U_2 appear "near" each other on a page, they are likely to have similar topics or some other shared characteristic. Nearness can be defined as proximity in the structural hierarchy of the page.*

Further context is given by the headers and text immediately preceding a list (such as "Family" and "Relatives with Web pages", respectively) and by the text preceding and anchoring a link (such as "My sister" and "Andrea", respectively).

Other Types of Links

Above, we discussed the hierarchies of hyperlinks, directories, and the structure of a page. Other links implicit or explicit in publicly-available information are:

Domain Names*: For example, inferring that "ellens@ai.mit.edu" and "erspert@athena.mit.edu" are both affiliated with the same educational institution. In the case of URLs, domain names can be considered part of the directory structure. For example, instead of categorizing the link from the homepage (www.ai.mit.edu/people/ellens/pers.html) to Tech Squares (www.mit.edu/activities/tech-squares/) as outward due to their being on different hosts, one could recognize that both sites are at "mit.edu". This might suggest that Tech Squares is more closely related to the individual in some way than is Puddletown Dancers (located at host www.glyphic.com).*

Relationships between concepts represented by words and phrases*: We can regard there as being a link between pieces of text that have related meanings. The simplest such relationship is between synonyms or multiple versions of a name (e.g., "Adina Dorothy Spertus Melhus" and "Adina Melhus"). More complex relations can be discovered through the the WordNet ontology [Miller 1995]; for example, it specifies that "kinswoman" is a specialization of "relatives" and that "sister" and "niece" are specializations of* *"kinswoman". This could lead to the inference that "my sister" is more similar to "my niece" than is "my brother".*

***Paths traveled through Web sites by visitors**, indicating a relation among the pages not revealed by the static links alone [Spertus 1995, Pirolli et al. 1996]: For example, if nearly everyone who followed the "Tech Squares" link also followed the "square dancing" link but not vice versa, the conclusion could be drawn that "Tech Squares" is of narrower interest than "square dancing" is.*

Applications

To demonstrate the value of web type and topic inference, I describe three applications that can make use of this information, which are in the process of being implemented as part of the ParaSite system.

Finding Moved Pages

Search engines frequently return obsolete URLs. In 1995, Selberg and Etzioni found that 14.9% of the URLs returned by popular search engines no longer point to accessible pages [Selberg and Etzioni 1995]. With the growth and aging of the web since their measurements, the percent of obsolete URLs returned may now be even higher. Currently, there are no utilities that try to track down moved pages.

Example problem

Consider the blurb and URL U_{old} returned by AltaVista in response to a search for "Cecilia Rodriguez":

> ***N California Rallies for Peace and Justice in Chiapas***
>
> *Northern California rallies for peace & justice in Chiapas by John Trinkl Alarm sparked quick action when Northern California activists learned of the... http://garnet.berkeley.edu:3333/.mags/NPW/.npw-0395/npw-0395-chiapas.html - size 6K - 24 Apr 95*

At the time the search was performed, U_{old} no longer existed. The user would like to automatically be referred to the page's new location. Here we describe two approaches, using link-based heuristics.

Approach 1: Exploiting hyperlinks

One approach to automatically finding the new URL U_{new} is based on the observation that the people most likely to know the new location of the page are those who cared enough about the material to point to the old location (in the past):

> ***Heuristic 8 (Temporal Locality of References):** If a page R referenced a page P in the past, it is a good place to look for a reference to P, even if the URL of P has changed.*

Accordingly, a tool could request from AltaVista the pages R that referenced U_{old} (using the "link:" designator) at the time they were last crawled. It could then check whether each page R pointed to the new location U_{new} of the moved page, either directly or recursively, preferentially expanding links whose anchor text included blurb terms, such as "Chiapas". It would return a page to the user if the page matched the criteria of the original search (e.g., contained "Cecilia Rodriguez") and contained the information appearing *in the original blurb.*

We could expand this approach to use political hierarchies by learning from Yahoo that Chiapas is a region in Mexico and including "Mexico" among the terms sought in the anchor text.

Approach 2: Exploiting directory links

Heuristic 5 (Authorship Location) and Heurstic 6 (Directory/Hyperlink Correlation) tell us that the homepage of the author of U_{old} is likely to be in the same directory as U_{old} or in a parent directory. Furthermore, this page is likely to point to the new URL U_{new}. Accordingly, we remove fields from the right side of U_{old} until we find a page P that still exists. We then crawl from R to seek U_{new}.

Finding Related Pages

A common technique for finding pages similar to a given page P is collaborative filtering [Shardanand and Maes 1995], where pages are recommended that were liked by other people who liked P. This is based on the assumption that items thought likeable/similar by one user are likely to by another user. As collaborative filtering is currently practiced, users explicitly rate pages to indicate their liking. Our Heuristic 7 (Spatial Locality of References) can be seen as an extension of collaborative filtering in which liking is indicated not through explicit ratings but by observing hyperlinks: If a person links to pages P and Q, we can guess that people who like P may like Q, especially if the links to P and Q appeared near each other on the referencing page (such as within the same list).

Accordingly, if a user requests a page similar to a set of pages {P_1, ... P_n}, ParaSite can find (through AltaVista) pages R that point to a maximal subset of these pages and then return to the user what other pages are referenced by R. Note that ParaSite does not have to understand what the pages have in common. It just needs to find a list that includes the pages and can infer that whatever trait they have in common is also exemplified by other pages they point to. For example, the first page returned from AltaVista that pointed to both Computer Professionals for Social Responsibility (www.cpsr.org) and Electronic Privacy Information Center (www.epic.org) was a list of organizations fighting the Communications Decency Act; links included Electronic Frontier Foundation and other related organizations.

A Person Finder

A brute force search for a person's full name and affiliation is often effective at finding his or her homepage, but sometimes other approaches are needed. For example, the page may not yet have been indexed, or the affiliation or full name might not be known. For example, we might only know that a woman with the first name Nell teaches CS at the University of Texas (UT) at Austin. Not having the full name, we couldn't perform a search directly. Instead, we would go to the UT Austin CS home page and crawl down links until finding the homepage of someone with first name "Nell". Such a search could be automated. If the specific school had not been known but only that it was in Texas, the crawl could start from a list of Texas universities intersected with a list of CS departments. Sometimes, only a last name and a profession is known. For example, it is common to know only the last name of an author in one's field. An automatic tool could support requests to find a computer scientist with last name "Crouch" by crawling from a list of computer science institutes.

If the person's full name is known, perhaps even more useful than searching for a page with the full name in its title would be searching for pages with the full name as anchor text and following the associated links.

While there is no single stereotypical title for homepages, there is for references to them. To take an extreme example, Nick Kushmerick's homepage is entitled "The Labyrinth of Mediocrity". An AltaVista search on ***anchor:"Nick Kushmerick"*** *returns two references, one of which still contains anchor text "Nick Kushmerick". Following that links leads to the desired homepage. This approach takes advantage of how the author of the referencing page decided to label the homepage. Note also that it takes advantage of the likelihood that the link that was found on the referencing page during the search engine's crawl is likely to still exist, as specified in Heuristic 8 (Temporal Locality of References).*

Conclusions

Related Work

Node and Link Type Inference

Typed links have long been a topic of interest, beginning with semantic net research [Woods 1985], and the Nanards have made strong arguments for typing anchors [Nanard and Nanard 1993]. In 1983, Trigg identified 80 different classes of links [Trigg 1983]. Numerous people have since advocated typed links [Shum 1996, Halasz 1991], but very little work has been done on inferring the types of unlabelled links. An exception is Allan's recent work on link type inference based on inter-document similarity and structural information, although that does not handle anchors, only links between the wholes of two documents [Allan 1996]. On the whole, hypertext research seems not to fully address the differences between the Web and classical hypertext, such as the qualitative novelty of inter-site links, due to deception and differing standards.

Numerous researchers, including Furuta [1994] and André et al. [1989] have written about structured documents and the benefits they provide for consistency, reusability, and verifiability, but links between documents have been less thoroughly explored. Pirolli et al. use inter-node topology for node type inference, although they limit themselves to a single site. To encapsulate the relationships among Web pages at a site, they maintain three graph structures. One indicates the existence of a link, the second inter-page text similarity, and the third the flow of users. They used this data to classify pages into one of 8 types (such as "individual home page"), with limited success.

Using Information in Hyperlinks

Boyan et al. have observed that Web pages differ from general text in that they possess external and internal structure [Boyan et al. 1996]. They use this information to propagate rewards from interesting pages to those that point to them (also done by LaMacchia [1996]) and to more heavily weight words in titles, headers, etc., when considering document/keyword similarity, a technique used earlier in Lycos by Mauldin and Leavitt [1994]. Mauldin has made use of link information by naming a page with the anchor text of hyperlinks to it. Iwazume et al. have preferentially expanded hyperlinks containing keywords relevant to the user's query [Iwazume et al.], although O'Leary observes that anchor text information can be unreliable [O'Leary 1996]. LaMacchia has implemented or proposed heuristics similar to some mentioned in this paper, such as making use of the information in directory hierarchies [LaMacchia 1996]. Frei and Stieger have discussed how knowledge of the adjacency of nodes via hyperlinks can be used to help a user navigate or find the answer to a query [Frei and Stieger 1992].

ParaSite is one of the few systems to take a unified approach to mining a variety of types of link information.

The most related system is WebSQL, which treats the Web as a relational database. While the published work is based primarily on the hyperlink relation [Mendelzon 1996], work in progress allows the definition of other types of links.

Summary

The World-Wide Web has important qualitative and quantitative differences from traditional hypertext, creating new information retrieval problems but also presenting useful new varieties of information, which are currently underutilized. We have shown that heuristics that make use of information in a variety of kinds of links can be useful. Planned future research will determine the accuracy of the heuristics under various conditions, the usefulness of the applications relying on them, and a toolkit for building such applications.

Acknowledgments

I have received valuable help from my supervisor Lynn Andrea Stein; from my readers Ken Haase, Tom Knight, and Pattie Maes; and from Oren Etzioni, Keith Golden, Shelley Heller, David Karger, Brian LaMacchia, Nate Osgood, and Jerry Sussman. The author is a graduate student at MIT and a visiting scholar at University of Washington.

References

Allan 1996
James Allan. Automatic Hypertext Link Typing. In Proceedings of the Seventh ACM Conference on Hypertext, Washington, D.C., 1996.

André et al. 1989
Jacques André, Richard Furuta, and Vincent Quint, editors. Structured Documents. Cambridge University Press, 1989.

Andrews 1996
Keith Andrews. Applying Hypermedia Research to the World Wide Web. Position paper for the workshop "Hypermedia Research and the World-Wide Web", Seventh ACM Conference on Hypertext, Washington, D.C., 1996.

Boyan et al. 1996
Justin Boyan, Dayne Freitag, and Thorsten Joachims. A Machine Learning Architecture for Optimizing Web Search Engines. In The AAAI-96 Workshop on Internet-based Information Systems.

Franz 1996
Alexander Franz, editor. The AAAI-96 Workshop on Internet-based Information Systems. AAAI, August 1996.

Frei 1995
H.P. Frei and D. Stieger. The Use of Semantic Links in Hypertext Information Retrieval.Inform. Proc. and Management, Vol. 31, No.1, 1995, pp. 1-13.

Furuta 1994
Richard Furuta. Defining and Using Structure in Digital Documents. In Proceedings of the First

Annual Conference on the Theory and Practice of Digital Libraries, College Station, TX, 1994.

Halasz 1991

Frank Halasz. "Seven Issues" Revisited(keynote address). In Proceedings of the Third ACM Conference on Hypertext, San Antonio, TX, 1991.

Iwazume et al. 1996

Michiaki Iwazume, Kengo Shirakami, Kazuaki Hatadani, Hideaki Takeda, and Toyoaki Nishida. IICA: An Ontology-based Internet Navigation System. In The AAAI-96 Workshop on Internet-based Information Systems, 1996.

LaMacchia 1996

Brian A. LaMacchia. Internet Fish. PhD thesis, MIT Department of Electrical Engineering and Computer Science, 1996.

Mauldin and Leavitt 1994

Michael Mauldinand John R. R. Leavitt. Web Agent Related Research at the Center for Machine Translation. In Proceedings of the ACM Special Interest Group on Networked Information Discovery and Retrieval, 1994.

Mendelzon et al. 1996

Alberto O. Mendelzon, George A. Mihaila, and Tova Milo. Querying the World Wide Web. Proceedings of the Fourth International Conference on Parallel and Distributed Information Systems, Miami, FL, 1996.

Miller 1995

George A. Miller. WordNet: A Lexical Database for English. Communications of the ACM, pages 39-41, November 1995. See also http://www.cogsci.princeton.edu/~wn/.

Nanard and Nanard 1993

Jocelyne Nanard and Marc Nanard. Should Anchors be Typed too? An experiment with MacWeb. In Proceedings of the Fifth ACM Conference on Hypertext, pages 51-62, Seattle, WA, 1993.

O'Leary 1996

Daniel E. O'Leary. The Relationship Between Relevance and Reliability in Internet-based Information and Retrieval Systems. In The AAAI-96 Workshop on Internet-based Information Systems, 1996.

Pirolli et al. 1996

Peter Pirolli, James Pitkow, and Ramana Rao. Silk from a Sow's Ear: Extracting Usable Structures from the Web. In CHI '96 Proceedings: Conference on Human Factors in Computing Systems: Common Ground, pages 118-125, Vancouver, BC, 1996.

Selberg and Etzioni 1995

Erik Selbergand Oren Etzioni. Multi-Service Search and Comparison using the MetaCrawler. Proceedings of the 4th International World Wide Web Conference, 1995.

Shardanand and Maes 1995

Upendra Shardanand and Pattie Maes. Social Information Filtering: Algorithms for Automating "Word of Mouth". CHI '95 Proceedings: Conference on Human Factors in Computing Systems: Mosaic of Creativity, 1995.

Shum 1996

Simon Buckingham Shum. The Missing Link: Hypermedia Usability Research & The Web. URL: http://kmi.open.ac.uk/~simonb/missing-link/ml-report.html.

Spertus 1995

Ellen Spertus. Information Hierarchies. In Fifth Annual MIT Student Workshop on Scalable Computing. MIT Laboratory for Computer Science, August 1995.

Spertus 1996

Ellen Spertus. Mining Links. Thesis proposal, MIT Department of Electrical Engineering and Computer Science, September 1996.

Spertus and Lauckhart 1996

Ellen Spertusand Greogry Lauckhart. Link Geometry and Crawling on Demand. In Distributed Indexing/Searching Workshop, World Wide Web Consortium, May 1996.

Trigg 1983

Randall H. Trigg. A network-based approach to text handling for the online scientific community. PhD Thesis, University of Maryland, 1983. Also technical report TR-1346.

Woods 1985

William A. Woods. "What's in a Link: Foundations for Semantic Networks". In Ronald J. Brachman and Hector Levesque, eds., Readings in Knowledge Representation. Morgan Kaufmann, 1985.