Chapter 5

Applications

In this chapter, I describe some applications, written in Squeal. These applications were originally proposed in my thesis proposal [42].

5.1 Sibs: Finding Similar Pages

One common technique for finding pages similar to those in a given set P is to search for pages with words that appear frequently in the pages of P [13]. Another approach is collaborative filtering, where pages are recommended that were liked by other people who liked those in P; whether a user liked a page is determined by their explicit rating [40]. My technique is a variant type of collaborative filtering where liking is indicated not through explicit ratings but by observing hyperlinks. Specifically, a page that points to elements of P is likely to point to similar pages.

5.1.1 Basic Technique

One could use the following algorithm to find pages similar to $P1$ and $P2$:

1. Generate a list of pages R that reference $P1$ and $P2$.
2. List the pages most commonly referenced by pages within R.

Figure 5-1 shows the Squeal code implementing this algorithm. Figure 5-2 shows a transcript.

5.1.2 Optimizations

Some optimizations/variations are:

1. Only return target pages that include a keyword specified by the user (Figure 5-3). Adding the keyword "women" removed the links to the ACLU and PFAW.
2. Return the names of hosts frequently containing referenced pages (Figure 5-4). This brought to prominence Electronic Policy Network (efn.org) and Close Up Foundation (www.closeup.org).
3. Only return target pages that point to one or both of $P1$ and $P2$ (Figure 5-5). The only such page (besides the NOW home page, which points to itself) was the English Server's page on Feminism and Women's Studies ("http://english-www.hss.cmu.edu/feminism.html"), which points to both.
4. Only follow links that appear in the same list (Section 2.2.3) and under the same header (Section 2.2.3) as the links to $P1$ and $P2$ (Figure 5-6). This causes all of the pages to drop substantially below the original pages. Remaining pages include the ACLU and PFAW.

```
DEFPROC SimPagesBasic(page1id, page2id, threshold)
    // Create temporary data structures
    CREATE TABLE parent(url_id url_id);
    CREATE TABLE results(url_id url_id, score INT);

    // Insert into "parent" the pages that reportedly link
    // to both pages that we care about
    INSERT INTO parent (url_id)
    SELECT DISTINCT r1.source_url_id
    FROM rlink r1, rlink r2
    WHERE r1.source_url_id = r2.source_url_id AND
    r1.dest_url_id = page1id AND
    r2.dest_url_id = page2id;

    // Store the pages pointed to by the parent pages,
    // along with a count of the number of links to them
    INSERT INTO results (url_id, score)
    SELECT l.dest_url_id, COUNT(*)
    FROM link l, parent p
    WHERE l.source_url_id = p.url_id
    GROUP BY l.dest_url_id;

    // Show the URLs of pages most often pointed to
    // and the number of links to them
    SELECT v.vcvalue, COUNT(*)
    FROM link l, parent p, valstring v, urls u
    WHERE l.source_url_id = p.url_id
    AND l.dest_url_id = u.url_id
    AND u.value_id = v.value_id
    GROUP BY v.vcvalue
    HAVING COUNT(*) >= threshold
    ORDER BY COUNT(*) DESC;
ENDPROC;
```

Figure 5-1: Code for SimPagesBasic

The home page for the National Organization for Women (NOW)
? url_id('www.now.org');
1877

The home page for the Feminist Majority Foundation
? url_id('www.feminist.org');
1503

SimPagesBasic(1877, 1503, 3);

vcvalue	
http://www.now.org/	13
http://www.feminist.org/	7
http://www.aauw.org/	4
http://www.aclu.org/	4
http://www.feminist.org/gateway/womenorg.html#top	4
http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mmbt/www/women/writers.html	3
http://www.cs.utk.edu/ bartley/other/ISA.html	3
http://www.democrats.org/	3
http://www.feminist.org/fmf/graphics/navigate.map	3
http://www.igc.org/igc/womensnet/	3
http://www.pfaw.org/	3

Figure 5-2: Transcript of run of SimPagesBasic. The variable **tolinks** (Section 3.6) is set to 20. Acronyms include AAUW (American Association of University Women) and PFAW (People for the American Way).

```
SELECT v.vcvalue, r.score
FROM results r, valstring v, urls u, page pg
WHERE u.url_id = r.url_id
AND v.value_id = u.value_id
AND pg.url_id = u.url_id
AND pg.contents LIKE strcat("%", keyword, "%")
AND r.score >= threshold
ORDER BY r.score DESC;
```

Figure 5-3: Modification to SimPagesBasic (Figure 5-1) to require presence of keyword.

```
SELECT par.value, COUNT(*)
FROM link l, parent p, urls u, parse par
WHERE l.source_url_id = p.url_id
AND l.dest_url_id = u.url_id
AND par.url_value_id = u.value_id
AND par.component = 'host'
GROUP BY par.value
HAVING COUNT(*) >= threshold
ORDER BY COUNT(*) DESC;
```

Figure 5-4: Follow-On to SimPagesBasic (Figure 5-1) to return hosts.

```
SELECT v.vcvalue, r.score
FROM results r, valstring v, urls u, link l
WHERE u.url_id = r.url_id
AND v.value_id = u.value_id
AND l.source_url_id = u.url_id
AND r.score >= threshold
AND (l.dest_url_id = page1id OR l.dest_url_id = page2id)
ORDER BY r.score DESC;
```

Figure 5-5: Follow-On to SimPagesBasic(Figure 5-1) to only return pages pointing to one or more of the original pages.

5.1.3 Evaluation

The text-based approach to recommender systems is used by Excite (www.excite.com), which allows the user to request pages textually similar to those in P. Because Excite can only find pages similar to a single page, not a set of them, we only provide a single URL to each system for each round of the test.

Method

Five subjects agreed to evaluate the system. Each submitted a list of between 10 and 15 URLs that interested him. (All subjects were male.) For each of the URLs, I performed an Excite and ParaSite search for similar pages. One submitted page was "http://www.suntimes.com/ebert/ebert.html", with title value "Roger Ebert on Movies". We will refer to this as the seed URL.

Excite ratings The Excite engine takes a query as input and returns a set of URLs, each with a relevance score, summary, and a link offering to find "more like this". Figure 5-7 shows the first five listings returned for a query on "Roger Ebert".

While the first URL returned from Excite is slightly different from the seed URL, the associated pages have the same contents, which makes them the same for Excite's purposes. In order to get Excite's suggestions for more pages "like" the first one, we follow the appropriate hyperlink. Figure 5-8 shows the top 5 URLs returned by this second query (not showing their summaries and "more like this" links). It is from this list that we obtain the Excite-generated URLs similar to the seed URL. We take the first items returned from this second stage, skipping broken links and pages with the same content as the seed URL and broken links.

Of the first 25 user-submitted URLs that I received, 17 of them could be found in the Excite database.

Parasite ratings The code for Sibs is in Figure 5-9. It only considers links occurring at the same header and list levels as links to the user-specified page. The threshold for the number of common links was set to 2. In our evaluation, the "tolinks" command-line variable was set to 20 and "maxpage" to 30k (section 3.6).

Table 5.1 shows the top 5 URLs yielded by each system for the "Roger Ebert" query.

Of the 17 seed URLs for which Excite was able to generate similar URLs, ParaSite was able to find similar URLs in 13 cases. Table 5.2 summarizes the performance of Excite and ParaSite on the 25 potential seed URLs.

There were thirteen seed URLs for which sufficient pages were generated by each system. These seeds and the recommended URLs were given to the five subjects, with a request to rate the relevance, interestingness, and novelty of each recommended page relative to the seed. The complete instructions are shown in Figure 5-10. Seed URLs and the associated recommendations were displayed as shown in Figure 5-11, with each URL represented as a hyperlink to the specified page and

```
DEFPROC SimPageListHeader(page1id, page2id, threshold)
    CREATE TABLE parent(url_id url_id, hstruct BINARY(6), lstruct BINARY(6));
    CREATE TABLE results(url_id url_id, score INT);

    // Only put pages into "parent" if the links to page1id and
    // page2id appear at the same levels in the list and header;
    // store the hierarchy information with the url_id.
    INSERT INTO parent (url_id, hstruct, lstruct)
    SELECT DISTINCT l1.source_url_id, l1.lstruct, l1.hstruct
    FROM link l1, link l2
    WHERE l1.source_url_id = l2.source_url_id AND
    l1.dest_url_id = page1id AND
    l2.dest_url_id = page2id AND
    l1.lstruct = l2.lstruct AND
    l1.hstruct = l2.hstruct;

    // Store the pages pointed to by the parent pages
    // at the appropriate levels, along with a count of the number
    // of qualifying links
    INSERT INTO results (url_id, score)
    SELECT l.dest_url_id, COUNT(*)
    FROM link l, parent p
    WHERE l.source_url_id = p.url_id
    AND l.hstruct = p.hstruct
    AND l.lstruct = p.struct
    GROUP BY l.dest_url_id;

    // Show the URLs of pages most often pointed to
    // and the number of links to them
    SELECT v.vcvalue, COUNT(*)
    FROM link l, parent p, valstring v, urls u
    WHERE l.source_url_id = p.url_id
    AND l.dest_url_id = u.url_id
    AND u.value_id = v.value_id
    GROUP BY v.vcvalue
    HAVING COUNT(*) >= threshold
    ORDER BY COUNT(*) DESC;
ENDPROC;
```

Figure 5-6: Code for SimPageListHeader only counts links at the same level in the header and list hierarchies as the input pages.

78% Roger Ebert on Movies
URL: http://www.suntimes.com/ebert/
Summary: Roger Ebert on Movies. Search Roger Ebert's Reviews Current Roger Ebert Reviews Roger Ebert Features One Minute Movie Reviews Roger Ebert's The Great Movies Movie Answer Man.
More Like This: Click here to perform a search for documents like this one.

78% Russ Meyer
URL: http://www.well.com/user/jeffdove/russmeyer.html
Summary: A classic piece from 1973 in which Ebert sums up Meyer's career to that point and reflects on his own involvment with Beyond the Valley Of the Dolls. A 1973 interview by Kenneth Turan and Stephen F. Zito with some good biographical background and philosophy on filmmaking information from Meyer.
More Like This: Click here to perform a search for documents like this one.

77% Roger Ebert's Book of Film
URL: http://www.wwnorton.com/catalog/fall96/ebert.htm
Summary: For this delicious, instructive, and vastly enjoyable anthology, Roger Ebert has selected and introduced an international treasury of more than 100 selections that touch on every aspect of filmmaking and filmgoing. Here is a book to get lost in and return to time and time again–at once a history, an anatomy, and a loving appreciation of the central art form of our time.
More Like This: Click here to perform a search for documents like this one.

77% Roger Ebert's Book of Film
URL: http://www.wwnorton.com:81/catalog/fall96/ebert.htm
Summary: For this delicious, instructive, and vastly enjoyable anthology, Roger Ebert has selected and introduced an international treasury of more than 100 selections that touch on every aspect of filmmaking and filmgoing. Here as well are the novelists who have indelibly captured the experience of moviegoing in our lives (Walker Percy, James Agee, Larry McMurtry) and the culture of the movie.
More Like This: Click here to perform a search for documents like this one.

75% Siskel and Ebert's Web Picks
URL: http://w3.arizona.edu/ uab/picks.htm
Summary: Big Book Yellow Pages To help find the theaters.
More Like This: Click here to perform a search for documents like this one.

Figure 5-7: The First Five Listings Returned by Excite Query on "Roger Ebert". All hyperlinks are underlined.

Excite	ParaSite
www.suntimes.com/show/index.html *Showcase*	movieweb.com/movie/movie.html *MOVIEWEB: Home Page*
www.suntimes.com/bruno/bruno.html *Restaurant Reviews*	www.hollywood.com/ *Hollywood Online*
www.suntimes.com/savage/savage.html *Terry Savage on Personal Finance*	us.imdb.com *The Internet Movie Database (IMDb)*
www.suntimes.com/index/ *Chicago Sun-Times Online*	mirrors.yahoo.com/eff/speech.html *Black Thursday*
www.girlsonfilm.com/ *Girls On Film*	movieweb.com/movie/top25.html *MOVIEWEB: Box Office Stats*

Table 5.1: Top 5 Pages Returned by Excite and ParaSite with Ebert Seed URL (www.suntimes.com/ebert/ebert.html). For each page, the URL and title are listed, the title in italics.

99% Roger Ebert on Movies
URL: http://www.suntimes.com/ebert/

98% The Roger Ebert Movie Files
URL: http://www.suntimes.com/ebert/ebertser.html

96% Chicago Sun-Times Online
URL: http://www.suntimes.com/index/

96% Showcase
URL: http://www.suntimes.com/show/index.html

95% Chicago Sun-Times Online
URL: http://www.suntimes.com/index/index.html

Figure 5-8: Top Five Listings Returned by Excite "More Like This" Query on Ebert Seed URL.

```
DEFPROC Sibs(pageid, threshold)
    CREATE TABLE parent(url_id url_id, hstruct BINARY(6), lstruct BINARY(6));

    INSERT INTO parent (url_id, hstruct, lstruct)
    SELECT DISTINCT source_url_id, hstruct, lstruct
    FROM link
    WHERE dest_url_id = pageid;

    SELECT v.vcvalue, COUNT(*)
    FROM link l, parent p, valstring v, urls u
    WHERE l.source_url_id = p.url_id
        AND l.dest_url_id = u.url_id
        AND u.value_id = v.value_id
        AND l.hstruct = p.hstruct
        AND l.lstruct = p.lstruct
    GROUP BY v.vcvalue
    HAVING COUNT(*) >= threshold
    ORDER BY COUNT(*) DESC;
ENDPROC
```

Figure 5-9: Code for Sibs

URL	Excite (%)		ParaSite (#)	
	max	min	max	num
www.weather.com/weather/us/cities/TX_Austin.html	98	98	5	5
www.suntimes.com/ebert/ebert.html	95	92	4	5
us.imdb.com/search	99	98	n/a	0
www.mapquest.com/	95	82	3	5
www.americanair.com:80/aa_home.htm	98	98	3	5
www.texasbbq.com	–	–		
www.geodesic.com	96	94	3	4
www.mos.org/leonardo	82	80	?	?
www.amd.com	93	90	7	5
www.odci.gov/cia/dst/	93	87	n/a	0
www.cnn.com	–	–		
banking.wellsfargo.com	–	–		
www.ebay.com	99	96	3	1
www.metacrawler.com	–	–		
www.roguemarket.com	81	71	2	5
www.artbell.com	98	96	4	2
www.wsdot.wa.gov/regions/northwest/NWFLOW/	–	–		
members.aol.com/gwattier/washjob.htm	–	–		
www.activision.com	97	95	3	5
www.happypuppy.com	92	90	2	4
www.economist.com	86	73	3	5
www.owlnet.rice.edu/ indigo/gsotd/may96.html	90	84	3	5
wombat.doc.ic.ac.uk/foldoc/index.html	–	–		
www.wolfram.com/graphics/	–	–		
www.cs.ubc.ca/nest/imager/contributions/scharein/KnotPlot.html	99	93	6	5

Table 5.2: Performance of Excite and ParaSite on 25 Seed URLs. The top 4 or 5 URLs were picked from those generated by each system, assuming enough were generated. Except when the URL wasn't in Excite's database, indicated by dashes, Excite always generated more than five URLs. The maximum and minimum relevance ratings of the top 5 are listed. ParaSite was only run on URLs in Excite's database. The "max" column indicates the score of the best result, and "num" indicates the number of links available for the evaluation (i.e., with score 2 or more), to a maximum of 5. The URLs that were successful seeds for both systems are listed in bold face.

The purpose of this evaluation is to compare the Web pages suggested by two different systems in response to a seed URL provided by the user. In other words, the user enters a URL (s)he likes or considers interesting, and the system returns suggestions of more URLs to consider.

Each page of the evaluation consists of a seed URL and up to five URLs generated by each of the two systems. Take a look. You should print each of these pages to record your ratings and comments. For each returned URL, you will rate how relevant, interesting, and novel it is relative to the seed URL on a scale of 0 (not at all) to 3 (very much). Here are short descriptions of each of the criteria:

- **relevance**: how closely related to the seed URL
- **interestingness**: how interesting to someone interested in the seed URL
- **novelty**: how likely it is that the user didn't already know how to find the information on the page (i.e., is the suggested URL novel?)

If you were the person to submit the seed URL, give the ratings from your point of view. If you were not the person to submit the seed URL, imagine why they found the seed page interesting and try to answer from their point of view. Write down (or type in) your assumptions. Each URL uses the title of the page as the anchor text. Be sure to follow each link, because the title may not be accurate and may not give enough information. It should not usually be necessary to follow further links. You should also record your subjective impressions on the different recommendations and the different "feel" of the two systems. Please log the time you spend on each page. The first page will probably take you the longest. After that, you should take about 15-20 minutes per page, half on the ratings and half on comments.

Please contact me with any questions by email (ellens@cs.washington.edu), work phone (685-4087), or home phone (882-8669). After you have completed the forms, we can arrange a time for you to drop them off and for me to pay you. Thank you for participating.

There are **13** pages.

Figure 5-10: Instructions for evaluation of Recommender Systems

with the title of the page as the anchor text.

Results

As subjects pointed out, a rating for novelty seemed not to be applicable when a page was entirely irrelevant. For this reason, when "averaging" ratings, novelty was treated as zero when relevance was zero. To capture the simultaneous importance of the three metrics, I also list the product of the averages. A summary over all the pages is shown in Table 5.3. The complete ratings appear in appendix D. When scores for each subject are listed, individuals are identified by a single letter.

On average, the Excite pages were judged more relevant (1.84 vs. 1.36) and interesting (1.63 vs. 1.47) than the ParaSite pages, while the ParaSite pages were judged more novel (1.32 vs. 1.12) and had a higher product (4.58 vs. 4.29). The results of each page's evaluation of can be divided into three cases: those where all of the ParaSite averages are higher (3), where the Excite averages are higher (4), and where the results are mixed (6).[1] I discuss two pages in each category, including the most extreme.

Roger Ebert The recommendations for the Roger Ebert seed URL were shown in Table 5.1. The subject evaluations are shown in Table 5.4. Four of the Excite URLs were other pages at the *Chicago Sun-Times* site, which tended to get low ratings for relevance, interestingness, and especially novelty. The final Excite reference, "Girls on Film", a film-related zine was rated more highly, as were

[1] In one case, KnotPlot, all of the Excite ratings were higher except for novelty, where ParaSite performed better by 1%. I classified this as an entire win for Excite.

Ebert	Excite	0.95	1.30	0.55	0.92
	ParaSite	1.55	1.35	1.00	3.00
	Δ	63%	4%	82%	225%
Austin	Excite	0.95	0.95	0.25	0.23
	ParaSite	1.40	1.85	1.40	5.84
	Δ	47%	95%	460%	2459%
Mapquest	Excite	2.20	1.85	1.35	6.49
	ParaSite	0.90	1.40	0.78	1.59
	Δ	-59%	-24%	-43%	-76%
American	Excite	1.95	1.60	0.95	3.09
	ParaSite	0.99	1.11	1.45	7.00
	Δ	-49%	-31%	53%	127%
Geodesic	Excite	2.25	1.94	0.44	1.90
	ParaSite	2.31	2.13	1.75	9.13
	Δ	3%	10%	300%	381%
AMD	Excite	1.30	1.35	1.35	4.31
	ParaSite	1.81	1.69	0.88	3.34
	Δ	39%	25%	-35%	-23%
Rogue	Excite	1.30	1.35	1.35	4.31
	ParaSite	1.13	1.37	1.50	4.44
	Δ	-13%	1%	11%	3%
Art Bell	Excite	2.25	2.00	0.50	2.25
	ParaSite	0.6	1.00	0.90	0.85
	Δ	-73%	-50%	80%	-62%
Activision	Excite	1.80	1.45	2.15	5.73
	ParaSite	1.73	2.07	1.65	6.79
	Δ	-4%	43%	-23%	19%
Happy Puppy	Excite	2.63	2.19	1.13	7.15
	ParaSite	1.09	1.14	1.10	3.85
	Δ	-58%	-48%	-2%	-46%
Economist	Excite	1.90	1.70	1.80	6.32
	ParaSite	1.00	1.00	0.75	3.41
	Δ	-47%	-41%	-58%	-46%
Geek	Excite	1.94	1.44	0.94	3.12
	ParaSite	1.71	1.83	2.13	7.12
	Δ	-12%	28%	127%	128%
KnotPlot	Excite	2.45	2.05	1.85	9.95
	ParaSite	1.30	1.05	1.87	3.13
	Δ	-47%	-49%	1%	-69%
Average	Excite	1.84	1.63	1.12	4.29
	ParaSite	1.35	1.46	1.32	4.58
	Δ	-27%	-10%	17%	7%

Table 5.3: Averages of Ratings by Seed URL. Ratings were from 0 to 3, with higher numbers better. Seed URLs are listed in Table 5.2.

Seed URL: Roger Ebert on Movies

System A	System B
Showcase	MOVIEWEB: Home Page
Restaurant Reviews	Hollywood Online
Terry Savage on Personal Finance	The Internet Movie Database (IMDb)
Chicago Sun-Times Online	Black Thursday
Girls on Film	MOVIEWEB: Box Office Stats

Figure 5-11: Sample Evaluation Page for Recommender Systems. Underlined text indicates hyperlinks.

P	K	W	S	Title	Average			prod-
r i n	r i n	r i n	r i n		r	i	n	uct
1 1 1	1 1 1	3 3 1	1 1 0	Showcase	1.5	1.5	0.75	1.69
0 0 0	0 1 1	2 3 1	1 1 0	Restaurant Reviews	0.75	1.25	0.25	0.23
0 0 0	0 1 1	1 3 1	1 1 0	Terry Savage on Personal Finance	0.5	1.25	0.25	0.16
0 0 0	0 1 0	3 3 1	0 0 0	Chicago Sun-Times Online	0.75	1	0.25	0.19
1 2 2	2 1 2	2 3 1	0 0 0	Girls on Film	1.25	1.5	1.25	2.34
				Excite averages	0.95	1.30	0.55	0.92
1 1 1	2 2 2	3 3 1	2 2 2	MOVIEWEB: Home Page	2	2	1.5	6.00
0 0 0	2 2 2	3 3 1	2 1 0	Hollywood Online	1.75	1.5	0.75	1.97
0 0 0	2 2 2	2 2 1	3 0 1	The Internet Movie Database	1.75	1	1	1.75
0 0 0	0 0 3	2 1 1	0 0 0	Black Thursday	0.5	0.25	0.25	0.03
1 2 2	1 2 2	3 2 1	2 2 1	MOVIEWEB: Box Office Stats	1.75	2	1.5	5.25
				ParaSite averages	1.55	1.35	1.00	3.00

Table 5.4: User Ratings of Ebert Recommendations. "P", "K", "W", and "S" are pseudonyms for the subjects. The subject who gave the seed is in bold face. The letters "r", "i", and "n" stand for "relevance", "interestingness", and "novelty", respectively.

four of ParaSite's recommendations, all film-related. The remaining ParaSite reference, to "Black Thursday", a Web censorship protest, was judged entirely irrelevant. A representative comment was:

> I think the reason the person [submitting the URL] found this seed page interesting was because they wanted movie reviews. System A [Excite] tended to give results that were more related to where the seed page existed, unlike System B [ParaSite] which gave me the impression that it actually read and interpreted the content. —K

Quantitatively, users found the ParaSite results more relevant (1.55 vs. .95) and novel (1 vs. .55) than the Excite results, and about equally interesting. The product of relevancy, interestingness, and novelty was much higher for ParaSite than Excite (3 vs. .92).

Austin weather The URLs returned by each system for the page entitled "The Weather Channel - Austin, TX" are shown in Table 5.5. Excite returned Weather Channel reports on other cities in Texas and Arkansas, while ParaSite generally returned information, not necessarily weather-related about Austin. The user ratings are shown in Table 5.6. Three of the users, including the one who submitted the seed URL, preferred the ParaSite listings. The fourth reviewer thought that the weather information returned by Excite was more relevant. As with the previous seed, one of the URLs returned by ParaSite was deemed almost entirely irrelevant. Quantitatively, the ParaSite pages were judged more relevant (1.4 vs. .95), interesting (1.8 vs. .95), and novel (1.37 vs. .25) than the Excite pages. The product of the ratings was much higher for ParaSite, 5.29 vs. .23.

Excite	ParaSite
/weather/us/cities/TX_Lamesa.html	www.intellicast.com/weather/aus/
The Weather Channel - Lamesa, TX	Austin, TX
/weather/us/cities/TX_Seminole.html	www.austin360.com
The Weather Channel - Seminole, TX	*Austin 360: THE city site for Austin*
/weather/us/cities/TX_Pecos.html	www.texasmonthly.com/
The Weather Channel - Pecos, TX	*TEXAS MONTHLY WWW RANCH*
/weather/us/cities/TX_Muleshoe.html	pilot.msu.edu/user/ander299/jokes.htm
The Weather Channel - Muleshoe, TX	*Jokes*
/weather/us/cities/AR_Hot_Springs.html	www.auschron.com/current/music.clubs/
The Weather Channel - Hot Springs, AR	*City Beat... 8 Nights of Austin Music*

Table 5.5: Top 5 Pages Returned by Excite and ParaSite with Austin Weather Seed URL (http://www.weather.com/weather/us/cities/TX_Austin.html). For each page, the URL and title are listed. For the Excite URLs, the host name was omitted above; it is "www.weather.com" in each case.

P	K	W	S	Title	Average			prod-
r i n	r i n	r i n	r i n		r	i	n	uct
0 0 0	1 1 0	3 3 1	0 0 0	TWC - Lamesa, TX	1	1	0.25	0.25
0 0 0	1 1 0	3 3 1	0 0 0	TWC - Seminole, TX	1	1	0.25	0.25
0 0 0	1 1 0	3 3 1	0 0 0	TWC - Pecos, TX	1	1	0.25	0.25
0 0 0	1 1 0	3 3 1	0 0 0	TWC - Muleshoe, TX	1	1	0.25	0.25
0 0 0	0 0 0	3 3 1	0 0 0	TWC - Hot Springs, AZ	0.75	0.75	0.25	0.14
				Excite averages	0.95	0.95	0.25	0.23
3 3 3	2 3 2	3 3 1	3 2 3	Untitled	2.75	2.75	2.25	17.02
2 3 2	2 2 2	1 1 1	3 2 2	Austin 360...	2	2	1.75	7.00
1 2 2	1 1 2	1 2 2	2 2 2	TEXAS MONTHLY...	1.25	1.75	2	4.38
0 2 2	0 0 3	1 2 1	0 2 0	Jokes	0.25	1.5	0.25	0.09
1 2 2	0 0 3	1 2 1	1 1 0	City Beat...	0.75	1.25	0.75	0.70
				ParaSite averages	1.40	1.85	1.40	5.84

Table 5.6: User Ratings of Austin Weather Recommendations. "The Weather Channel" is abbreivated "TWC".

Excite	ParaSite
www.amd.com/K6/misc/articles.html *AMD-K6(TM) Related Articles*	www.intel.com/ *Welcome to Intel*
www.lubbockonline.com/news/040297/amd.htm *AMD unveils new K6 chip to rival Intel*	www.apple.com/ *Apple Computer*
www.pcworld.co.nz/pcwtop10/pcwtop10.shtml *@IDG New Zealand*	www.att.com/ *AT&T Home Page*
www.byte.com/art/9611/sec6/ART9.HTM *November 1996 / State Of The Art /* *The x86 Gets Faster with Age*	http://www.baynetworks.com/ Welcome to Bay Networks

Table 5.7: Top 4 Pages Returned by Excite and ParaSite for AMD Seed URL (www.amd.com). Because one of Excite's suggestions turned out to be in a foreign language, only the top 4 suggestions from each were used.

P	**K**	W	S	Title	Average			prod-
r i n	r i n	r i n	r i n		r	i	n	uct
3 3 0	2 2 0	3 3 1	2 1 0	AMD-K6(TM) Related Articles	2.5	2.25	0.25	1.41
3 2 2	2 1 0	3 3 1	2 1 0	AMD unveils ...	2.5	1.75	0.75	3.28
3 3 2	1 1 3	3 3 3	2 1 1	IDG New Zealand	2.25	2	2.25	10.13
3 0 1	1 1 3	3 3 3	2 3 1	November 1996 ...	2.25	1.75	2	7.88
				Excite averages	2.38	1.94	1.31	5.67
3 3 1	3 3 0	3 3 1	3 2 2	Welcome to Intel	3	2.75	1	8.25
3 3 1	2 1 0	1 1 1	1 2 2	Apple Computer	1.75	1.75	1	3.06
2 2 1	1 1 0	0 1 1	2 1 1	AT&T Home Page	1.25	1.25	0.5	0.78
2 2 2	0 0 3	1 1 1	2 1 1	Welcome to Bay Networks	1.25	1	1	1.25
				ParaSite averages	1.81	1.69	0.88	3.34

Table 5.8: User Ratings of AMD Recommendations

AMD The URLs returned by each system for the "Advanced Micro Devices" (AMD) home page are shown in Table 5.7. Excite generated URLs for pages with information specific to AMD, as the titles suggest. ParaSite generalized differently, returning the home pages of other computer companies. As one user wrote:

> It depends on the orientation of the originator
> is s/he interested in the company, in chips, in investment, in employment or what? —P

Three out of the four users considered the Excite results superior, giving it a higher score on all metrics, especially relevance and novelty, as shown in Table 5.8.

Geek Site of the Day

The URLs returned by each system for the "Geek Site of the Day" (GSotD) are shown in Table 5.9. Because ParaSite only made four recommendations, only the top four Excite recommendations are listed. Two of the Excite recommendations were articles about GSotD, one reviewed GSotD and similar sites, and one was a GSotD archive. The ParaSite selections were more diverse: the first two were collections of cool/useless pages, the next was the home page of "CNET: The Computer Network", and the fourth was the Museum of Bad Art. User ratings are shown in Table 5.10. The Excite pages were considered more relevant (1.94 vs. 1.71), while the ParaSite pages were considered more interesting (1.83 vs. 1.44) and novel (2.13 vs. 0.94). Users disagreed as to which system was preferable:

Excite	ParaSite
www.webcrawler.com./News/site.06.html *WebCrawler*	cool.infi.net *Cool Site of the Day*
www.owlnet.rice.edu/ indigo/gsotd/pcnovice.html *PC Novice mentions GSOTD*	www.go2net.com/internet/useless/ *go2net | internet | The Useless Pages*
www.owlnet.rice.edu/~indigo/gsotd/sept95.html *Geek Sites of the Day, September 1995*	www.cnet.com *Welcome to CNET.COM*
www.newsherald.com/BUSINESS/B20.HTM *News Herald Business Wire: How to keep up with everything that's cool*	www.glyphs.com/moba/ *The Museum of Bad Art*

Table 5.9: Top 4 Pages Returned by Excite and ParaSite for GSotD Seed URL (www.owlnet.rice.edu/~indigo/gsotd/).

P	K	W	**S**	Title	Average			prod-
r i n	r i n	r i n	r i n		r	i	n	uct
1 1 1	2 1 0	3 3 3	3 2 1	WebCrawler	2.25	1.75	1.25	4.92
2 1 0	2 1 0	1 1 0	1 0 0	PC Novice mentions GSOTD	1.5	0.75	0	0.00
3 3 0	2 1 0	3 3 3	1 0 0	GSotD, September 1995	2.25	1.75	0.75	2.95
3 2 3	1 1 3	0 1 1	3 2 1	News Herald Business Wire	1.75	1.5	1.75	4.59
				Excite averages	1.94	1.44	0.94	3.12
3 3 3	2 2 2	0 1 2	3 1 2	Cool Site of the Day	2	2.25	2	9.00
3 3 3	2 2 2	2 2 3	3 3 3	go2net: The Useless Pages	2.08	2.08	2.5	10.85
3 3 3	0 0 1	1 1 3	- - -	Welcome to CNET.COM	1.75	1.75	2	6.13
1 2 3	1 1 3	2 2 2	3 3 2	The Museum of Bad Art	1	1.25	2	2.50
				ParaSite averages	1.71	1.83	2.13	7.12

Table 5.10: User Ratings of Geek Site of the Day Recommendations. Dashes indicate where a user neglected to rate a page.

> System A [Excite] came up with one good suggestion. System B [ParaSite] came up with several. System B wins... —P
>
> I assume that the person wants sites that would be interesting or funny to the computer geek, such as things in poor taste. In this case I would choose system A. —W

MapQuest The URLs returned by each system for the "MapQuest!" home page are shown in Table 5.11. All 5 sites returned by Excite were highly-relevant map-related sites. The 5 sites returned by ParaSite were all related to travel but much less directly, such as tourist information about San Diego. The user ratings are shown in Table 5.12. Excite was rated better for all measures.

MapQuest The KnotPlot Site contains pictures of mathematical knots and links generated by a program called KnotPlot by its author Rob Scharein. The URLs returned by each system for the KnotPlot page are shown in Table 5.13. Three of the sites returned by Excite were knot-related sites on the same machine; one other was an announcement of a talk Scharein gave on KnotPlot; and the fifth was the home page of a topologist with a link to the KnotPlot Site. Of the five ParaSite suggestions, only one was of similar quality: Rob Scharein's home page. Three of the remainder were math-related pages, and one was the home page of an individual with no apparent relevance. The user ratings are shown in Table 5.14. The two systems got approximately the same score for novelty, and the Excite pages were judged significantly more relevant and interesting.

Excite	ParaSite
www.mckinley.com/magellan/Reviews/ News_and_Reference/Geography_and_Maps/ index.magellan.html *Geography & Maps Topics*	www.lib.utexas.edu/Libs/PCL/Map_ collection/Map_collection.html *PCL Map Collection*
netfind.aol.com/aol/Reviews/News_and_ Reference/Geography_and_Maps/ index.netfind.html *AOL NetFind: Reviews: Geography & Maps*	metro.jussieu.fr:10001/bin/cities/english *Subway navigator*
www.yumasun.com/-feat/Net/netmaps.html *Maps on Net*	pubweb.parc.xerox.com/map *Xerox PARC Map Viewer*
www.geosys.com/ *GeoSystems: Welcome!*	www.geocities.com/HotSprings/4150 *Inside San Diego*
elmo.webcrawler.com/select/trref.24.html *MapQuest*	www.geocities.com/HotSprings/4150/info.html *Inside San Diego: INFORMATION REQUEST*

Table 5.11: Top 5 Pages Returned by Excite and ParaSite for MapQuest Seed URL (http://www.mapquest.com).

P	K	W	**S**	Title	Average			prod-
r i n	r i n	r i n	r i n		r	i	n	uct
2 2 3	2 2 2	3 3 1	3 2 1	Geography & Maps Topics	2.5	2.25	1.75	9.84
2 3 2	2 2 2	3 3 1	3 2 2	AOL NetFind...	2.5	2.5	1.75	10.94
1 2 1	2 2 2	3 2 1	3 2 2	Maps on Net	2.25	2	1.5	6.75
1 2 3	1 1 1	3 2 1	3 1 1	GeoSystems: Welcome!	2	1.5	1.5	4.50
0 0 0	2 0 0	3 3 1	2 1 0	MapQuest	1.75	1	0.25	0.44
				Excite averages	2.20	1.85	1.35	6.49
0 1 2	1 1 3	3 3 2	3 2 0	PCL Map Collection	1.75	1.75	1.125	3.45
1 1 1	0 0 3	2 3 2	1 3 2	Subway navigator	1	1.75	1.25	2.19
0 1 1	0 0 3	3 3 1	2 3 3	Xerox PARC Map Viewer	1.25	1.75	1	2.19
0 1 1	0 0 3	1 2 1	0 1 1	Inside San Diego	0.25	1	0.25	0.06
0 1 1	0 0 3	1 2 1	0 0 0	Inside San Diego: INFO...	0.25	0.75	0.25	0.05
				ParaSite averages	0.90	1.40	0.78	1.59

Table 5.12: User Ratings of MapQuest Recommendations

Excite	ParaSite
www.cs.ubc.ca/nest/imager/contributions/scharein/KnotSquare.html *The Knot Square*	www.cs.ubc.ca/spider/scharein *Rob Scharein's Main WWW Page*
www.cs.ubc.ca/nest/imager/contributions/scharein/knot-theory/fox-knot.html *A Fox's Quick Introduction to Knot Theory*	www.hk.super.net/~cismath *CIS Mathematics Department*
www.forum.swarthmore.edu/news.archives/geometry.announcements/article212.html *Untitled*	forum.swarthmore.edu/ *The Math Forum Home*
www.cs.ubc.ca/nest/imager/contributions/scharein/knot-theory/references.html *Excellent References on Knot Theory*	forum.swarthmore.edu/maw/ *Math Awareness Week 1997*
www.ma.utexas.edu/~sedgwick/ Eric Sedgwick	users.aol.com/wkrulac/bill.htm Bill Krulac's Home Page

Table 5.13: Top 5 Pages Returned by Excite and ParaSite for KnotPlot Seed URL (www.cs.ubc.ca/nest/imager/contributions/scharein/KnotPlot.html)

P	K	W	**S**	Title	Average			prod-
r i n	r i n	r i n	r i n		r	i	n	uct
3 3 3	3 3 1	3 3 2	1 0 0	The Knot Square	2.5	2.25	1.5	8.44
3 3 3	2 3 1	3 3 2	3 3 2	A Fox's Quick Introduction...	2.75	3	2	16.50
3 1 2	2 1 3	2 2 1	1 0 0	Untitled	2	1	1.5	3.00
3 3 3	2 1 3	3 3 2	1 0 0	Excellent References...	2.25	1.75	2	7.88
3 3 3	2 1 3	3 3 2	3 2 1	Eric Sedgwick	2.75	2.25	2.25	13.92
				Excite averages	2.45	2.05	1.85	9.95
- - -	1 0 3	3 3 2	2 2 2	Rob Scharein's... Page	2	1.5	2.33	7.00
1 1 3	1 0 3	1 1 1	2 1 1	CIS Mathematics Department	1.25	0.75	2	1.88
2 2 2	1 1 3	2 3 1	1 1 1	The Math Forum Home Page	1.5	1.75	1.75	4.59
1 1 2	1 0 3	2 2 2	1 0 1	Math Awareness Week 1997	1.25	0.75	2	1.88
1 1 3	0 0 3	1 1 2	0 0 0	Bill Krulac's Home Page	0.5	0.5	1.25	0.31
				ParaSite averages	1.30	1.05	1.87	3.13

Table 5.14: User Ratings of KnotPlot Recommendations

URL	count	description
www.cs.ubc.ca/spider/scharein	6	Rob Scharein's home page
www.geom.umn.edu/	4	The Geometry Center
www.geom.umn.edu/~scharein/knotplot/		KnotPlot-related pages...
resource/www/CommandWindow.html	4	
resource/www/ControlPanel.html	4	
resource/www/KnotPlotManual.html	4	
resource/www/KPManOverView.html	4	
resource/www/ViewWindow.html	4	
www.math.uiowa.edu/knotplot/		KnotPlot mirror pages...
CommandWindow.html	4	
ControlPanel.html	4	
KnotPlotManual.html	4	
KPManOverView.html	4	
ViewWindow.html	4	
www.metacreations.com/	4	A computer visualization company
www.earlham.edu/suber/knotlink.htm	3	A different knots page
www.unitedmedia.com/comics/dilbert/	3	The Dilbert comic strip

Table 5.15: Pages Returned by ParaSite for KnotPlot with tolinks=40

URL	count	description
www.yahoo.com/	6	Yahoo!
www.cmpnet.com/	3	CMPnet: The Technology Network
www.geocities.com/HotSprings/4150/x.html	3	[broken link]
www.mckinley.com/	3	Magellan Internet Guide
www.netguide.com/	3	NetGuide: Your Guide to the Net
www.usps.gov/ncsc/	3	USPS zip code information

Table 5.16: Pages Returned by ParaSite for MapQuest with tolinks=40

Discussion

The ParaSite suggestions were judged more novel, while the Excite ratings were judged more relevant and interesting. In some cases, one system was markedly superior to the other. Some possible conclusions are:

1. The text-based approach is likelier than the structure-based approach to stay within the seed web site, yielding pages that users find more relevant but less novel.
2. Neither of the two approaches is always superior. Whether the text- or structure-based approach is better depends on the type of link and the user's purpose.
3. A superior system could be built by combining the two approaches.
4. The structure-based approach would have generated more useful results if more pages had been examined for each seed URL.

To test the last hypothesis, I reran the tests for "KnotPlot" and "MapQuest" with **tolinks** doubled to 40, meaning that the system examined up to forty pages pointing to the seed URLs. The KnotPlot run took an hour. Table 5.15 shows the URLs with ratings of three or greater. Because the subjects were no longer available to me, I have no ratings for the recommendations, but they appear to be much better than the original ParaSite recommendations. The MapQuest run took two-and-a-half hours. The results, shown in Table 5.16, do not appear to be much better. This suggests that, if enough pages are examined, the structure-based approach is superior to the text-based approach in most but not all cases. Further testing is desirable.

Related work

Collaborative filtering [40] is based on the assumption (also underlying parts of this work) that some people have the same taste as others. Consequently, if two people have some judgments in common, something liked by the first person should be recommended to the second person. The similar page finder can be seen as an application of collaborative filtering. While collaborative filtering systems have required users to explicitly rate items, there is no reason implicit ratings could not be used, as is the case with data mining [9].

Also related is bibliometrics, the statistical study of documents, which includes citation indexing [37]. Binary relations studied include *bibliographic reference*, where one paper references another; *bibliographic coupling*, where two papers share a common reference [17]; and *co-citation*, where two papers are referenced by a common source [41]. Co-citation is equivalent to ParaSite's judging two web documents similar if they are both referenced by the same page. As with links among web pages, it has been observed that there are many different motivations behind citations, including people's tendency to cite their own papers and to engage in quid pro quos [37]. The term "sitation" has been coined by Gerry McKiernan to describe the study of links between Web pages [35].

5.2 A Home Page Finder

In Section 1.1.3, heuristics for a program to find personal home pages were discussed. The primary trick was to look at hyperlinks to find pages that others have labeled as being a specific person's home page; i.e., to look for pages that are the destination of hyperlinks whose anchor text contains the person's name. Further heuristics take advantage of the tag structure within a page and the structure of candidate URLs. We can thus build an application to find home pages given a person's name:

1. Consider a page to be a candidate if it is the destination of a hyperlink with anchor text containing the person's name (and, ideally, nothing else).
2. Give a bonus to candidate pages with any of these characteristics:
   - A tag attribute contains the full name.
   - The full name appears within title or header tags.
   - The URL is of the form "<foo>/<foo>.html".
   - The URL file name is "index.html" or "home.html" or the empty string.
   - The final directory name in the URL starts with a tilde ( ) (the Unix convention for a user's home directory).
   - The penultimate directory name in the URL is "people" or begins with "home".
3. Display pages with the highest scores, allowing the user to find out the system's reasons.

The top-level code for the home page finder is shown in Figure 5-12. The rest of the code appears in Appendix C.

5.2.1 Sample Run

Here is a transcript of a run of the home page finder and subsequent queries:

HomePage('Pattie Maes');

```
DEFPROC HomePage(fullname)
    CREATE TABLE possibleParent(url_id url_id, urlstring VARCHAR(255), reason CHAR(255));
    CREATE TABLE candidate(url_id url_id, score INT, reason CHAR(255));
    CREATE TABLE results(url_id url_id, score INT);

    HomePageCore(fullname);
    // Fill in the results table
    INSERT INTO results (url_id, score)
    SELECT c.url_id, SUM(c.score) AS tot
    FROM candidate c
    GROUP BY c.url_id;

    // Display the results
    SELECT u.url_id, v.vcvalue, r.score
    FROM valstring v, results r, urls u
    WHERE u.url_id = r.url_id
    AND u.variant = 1
    AND v.value_id = u.value_id
    ORDER BY r.score DESC;
ENDPROC;
```

Figure 5-12: Top level code for home page finder. HomePageCore is defined in Appendix C.

url_id	vcvalue	score
4279	http://pattie.www.media.mit.edu/people/pattie/	33
4539	http://lcs.www.media.mit.edu/people/pattie/	23
4315	http://www.media.mit.edu/ pattie	16
4395	http://altavista.digital.com/cgi-bin/query?pg=aq&what=web &fmt=.&q=Pattie+Maes%0D%0A%0D%0A%0D%0A &r=Pattie+Maes&d0=&d1=	3
4380	http://www-pcd.stanford.edu/pcd-archives/ pcd-seminar/1993-1994/0012.html	1
4392	http://delegate.tokai-ic.or.jp:18080/InfoServ/ Artec/artec017.htm	1
4574	ahttp://www.vpro.nl/www/arteria/maxk/maxk-dop23.html	1
4584	http://lcs.www.media.mit.edu/people/lieber/ Teaching/Int-Int/Int-Int-Announcement.html	1

What are the scoring reasons behind the top page?

SELECT score, reason FROM candidate WHERE url_id = 4279 ORDER BY score DESC;

score	reason
10	File name is " "
10	Penultimate directory is: "people"
2	Anchor from "http://lcs.www.media.mit.edu/groups/agents/papers.html" is the full name: "Pattie Maes"
2	Anchor from "http://nif.www.media.mit.edu/" is the full name: "Pattie Maes"
2	Anchor from "http://www-white.media.mit.edu/vismod/demos/smartroom/contributors/contrib.html" is the full name: "Pattie Maes"
2	Anchor from "http://www-yano.is.tokushima-u.ac.jp/research/moo/ismoo-e.html" is the full name: "Pattie Maes"
1	Anchor from "http://aries.www.media.mit.edu/people/aries/home-page.html" includes the full name: "Prof. Pattie Maes"
1	Anchor from "http://lcs.www.media.mit.edu/groups/agents/papers.html" includes the full name: "Pattie Maes"
1	Anchor from "http://nif.www.media.mit.edu/" includes the full name: "Pattie Maes"
1	Anchor from "http://www-white.media.mit.edu/vismod/demos/smartroom/contributors/contrib.html" includes the full name: "Pattie Maes"
1	Anchor from "http://www-yano.is.tokushima-u.ac.jp/research/moo/ismoo-e.html" includes the full name: "Pattie Maes"

Give the reasons for the scores of the pages below 5.

SELECT c.url_id, c.score, c.reason
FROM candidate c, results r
WHERE r.score < 5 **AND c.url_id = r.url_id;**

url_id	score	reason
4380	1	Anchor from "http://www-pcd.stanford.edu/pcd-archives/pcd-seminar/1993-1994/0013.html" includes the full name: "Terry Winograd: "PCD 1/14: Pattie Maes, MIT, Learning Interface Agents""
4392	1	Anchor from "http://www.mediamatic.nl/whoiswho/Maes/PattieMaes.html" includes the full name: "Pattie Maes and Bruce Blumberg"
4395	2	Anchor from "http://www.mediamatic.nl/whoiswho/Maes/PattieMaes.html" is the full name: "Pattie Maes"
4395	1	Anchor from "http://www.mediamatic.nl/whoiswho/Maes/PattieMaes.html" includes the full name: "Pattie Maes"
4574	1	Anchor from "http://www.vpro.nl/htbin/scan/www/arteria/maxk/maxk-dop07.html" includes the full name: "061194: Pattie Maes"
4584	1	Anchor from "http://lcs.www.media.mit.edu/people/lieber/Teaching/Teaching.html" includes the full name: "Intelligent Interfaces Seminar <I>Pattie Maes and Henry Lieberman</I>"

The system didn't recognize that the top three pages returned were identical, because their servers have different names. Let's cause it to sum the scores of pages based on their MD5 hash value.

SELECT p.md5, SUM(r.score) as score
FROM page p, results r
WHERE p.url_id = r.url_id
GROUP BY p.md5;

md5	score
1c4697007709d157ca508d500edde0bc	72
75715d6305d48582c577ccbb7ce0a25a	1
8c6a9679daede982bda0af2a8a9023ae	1
9cf56d81c2a3be5368f8126f9721aadf	1
aa3c0dc8ebdd3409f46d5d5113eeb3dc	3
ca9cca3ec2c4acdf199906f9611ae28f	1

Show the totals by url_id.

```
SELECT r.url_id,
      (SELECT sum(r2.score) FROM
      results r2, page p2
      WHERE r2.url_id=p2.url_id AND p2.md5=p.md5) AS cnt
FROM results r, page p
WHERE r.url_id = p.url_id
AND r.url_id =
(SELECT MIN(r3.url_id) FROM results r3, page p3 WHERE p3.url_id =
   r3.url_id AND p3.md5 = p.md5);
```

url_id	score
4584	1
4574	1
4395	3
4392	1
4380	1
4279	72

Tell me which of the pages contains "Lieberman".

```
SELECT r.url_id
FROM results r, page p
WHERE p.url_id = r.url_id AND p.contents LIKE '%Lieberman%';
```

url_id
4584

5.2.2 Support for nicknames

A problem with the above home page finder is that it uses a single string to represent a person's name. A human being would know that "Ken Haase" and "Kenneth Haase" might be the same person, even though the strings are different. To implement a home page finder that can make use of different versions of the same name, we create the table **names**, shown in Figure 5-13, and the code in Figure 5-14.

5.2.3 Evaluation

Names were taken from David Aha's list of Machine Learning and Case-Based Reasoning Home Pages (http://www.aic.nrl.navy.mil:80/ aha/people.html), which was chosen because it was used in the Ahoy! study [39]. The home page finder without nicknames was used (Figure 5-12), and the Squeal interpreter was prevented from making use of Aha's list or its mirrors. Of the first fifteen hyperlinks appearing under the letter "A", twelve links still worked. We also excluded the link for "David Aha", considering him a special case. Of the eleven remaining names, ParaSite successfully found nine home pages (82%) and failed in two cases (18%), as shown in Table 5.17. In one successful case, ParaSite found a better link than the one on Aha's list! Aha's entry for Lloyd Allison pointed to a one-page profile (http://www.cs.monash.edu.au/people/profiles/lloyd.html), while ParaSite returned a link to a more traditional home page (entitled "Lloyd Anderson - Home Page") that contained roughly two dozen links (http://www.cs.monash.edu.au:80/ lloyd/index.html). ParaSite unequivocally failed in

number	name
1	Ken
1	Kenneth
1	Kenny
1	Kennie
2	Deborah
2	Debby
2	Deb
2	Debbie
3	Thomas
3	Tom
3	Tommy

Figure 5-13: The "names" Table

```
DEFPROC HomePageWithNicknames(firstName, lastName)
    CREATE TABLE possibleParent(url_id url_id, urlstring VARCHAR(255), reason CHAR(255));
    CREATE TABLE candidate(url_id url_id, score INT, reason CHAR(255));
    CREATE TABLE results(url_id url_id, score INT);

    LET nameNumber = (SELECT number FROM names WHERE name=firstName);

    FETCH HomePageCore(fullname=strcat(n.name, ' ', lastName))
    FROM names n
    WHERE n.number = nameNumber;

    // Fill in the results table
    INSERT INTO results (url_id, score)
    SELECT c.url_id, SUM(c.score) AS tot
    FROM candidate c
    GROUP BY c.url_id;

    // Display the results
    SELECT u.url_id, v.vcvalue, r.score
    FROM valstring v, results r, urls u
    WHERE u.url_id = r.url_id
    AND u.variant = 1
    AND v.value_id = u.value_id
    ORDER BY r.score DESC;
ENDPROC;
```

Figure 5-14: Top Level Code for HomePageWithNicknames. HomePageCore is defined in Appendix C.

name	# returned	# correct	notes
Agnar Aamodt	2	2	separate pages for English, Norwegian
Gennady Argre	0	0	no anchors found
Kamal Ali	1	1	
Carolyn Allex	3	3	all equivalent
Lloyd Allison	5	1	better match than Aha
Ethem Alpaydin	1	1	
Rick Alterman	2	1	
Klaus-Dieter Althoff	3	2 or 3	
Tim Andersen	1	1	not counting two bad links returned
Bill Anderson	3	0	other Bill Andersons
Chuck Anderson	4	1	other Chuck Andersons

Table 5.17: Results of Home Page Finder on Names from Aha's List. Pages were only considered correct if they pointed to the *home page* of the specified individual.

Lenore Blum
Lenore Blum 1943- Written by Lisa Hayes, Class of 1998 (Agnes Scott College) Lenore Blum was a bright and artistic child who loved math, art, and music from her original introductions to them. Blum finished high school at the age of 16, after which...
http://www.scottlan.edu/lriddle/women/BLUM.HTM, 5359 bytes, 27Apr97

Figure 5-15: Blurb returned from HotBot in response to the query "Lenore Blum 1943"

its search for Bill Anderson, returning pages relevant to other people with the name. In the case of Chuck Anderson, which I considered a win, links were returned both to the intended Chuck Anderson and to others, clearly a hazard of only using names for searches.

5.3 Bo Peep: 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 [38]. 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.

5.3.1 Technique 1: Climbing the directory hierarchy

Figure 5-15 shows an AltaVista blurb containing a URL U_{bad} that is no longer valid. I describe a heuristic algorithm and application for finding a new URL U_{new} for the page, given U_{bad} and the title *title* of the page. In this example, U_{bad} ="http://www.scottlan.edu/lriddle/women/BLUM.HTM", and *title* = "Lenore Blum". (While the examples in this chapter involve topics that interest me, in no case are pages under my control involved in the results.)

We can create URL U_{base} by removing directory levels from U_{bad} until we obtain a valid URL. We can then crawl from U_{base} in search of a page with title *title*. This is based on the heuristic that someone who cared enough about the page to house it in the past is likely to at least link to the page now. Figure 5-16 shows code that creates a "candidate" data structure and seeds it with the parent of U_{bad}. Figure 5-17 shows how the **parse** and **valstring** tables might appear after the procedure is executed.

Figure 5-18 shows the complete program. The strategy is:

1. Let round = 0;
2. Insert the parent URL (as possible_url_id) and 0 (as round) into the **candidate** table.
3. Repeat:

```
DEFPROC bopeep_init(old_url_value_id)
    // Create a structure for candidates
    CREATE TABLE candidate (possible_url_id url_id, round INT);

    // Get an unused value_id
    LET new_value_id = value_id();

    // Copy all of the parse elements except for the file name
    // from the old_url_value_id to new_value_id
    INSERT INTO parse (url_value_id, component, value, depth)
    SELECT new_value_id, component, value, depth
    FROM parse
    WHERE url_value_id = old_url_value_id AND depth <> 1;

    // Set a null file name in the parse table for new_value_id
    INSERT INTO parse (url_value_id, component, value, depth)
    VALUES (new_value_id, 'path', '', 1);

    // Put the new url_id in the candidate table
    LET new_url_id = (SELECT url_id FROM urls WHERE value_id = new_value_id);
    INSERT INTO candidate (possible_url_id, round)
    VALUES (new_url_id, 0);
ENDPROC;
```

Figure 5-16: Bo Peep: Code to climb one up in the directory hierarchy

parse

url_value_id	component	value	depth
1	host	www.scottlan.edu	0
1	port	80	0
1	path	BLUM.HTM	1
1	path	women	2
1	path	lriddle	3
2	host	www.scottlan.edu	0
2	port	80	0
2	path		1
2	path	women	2
2	path	lriddle	3

valstring

value_id	vcvalue
1	http://www.scottlan.edu/lriddle/women/BLUM.HTM
2	http://www.scottlan.edu/lriddle/women/

Figure 5-17: Views of **parse** and **valstring** for a parent directory and child file

```
DEFPROC bopeep(old_url_value_id, title)
    bopeep_init(old_url_value_id);
    bopeep_loop(0, title);
ENDPROC;

DEFPROC bopeep_loop(round, title)

    // Check candidate table for matches
    SELECT DISTINCT c.possible_url_id
    FROM candidate c, tag t, att a, valstring v
    WHERE c.round = round AND
    t.url_id = c.possible_url_id AND
    t.name = 'title' AND
    a.tag_id = t.tag_id AND
    a.name = 'anchor' AND
    v.value_id = a.value_id AND
    v.vcvalue = title

    // Put children of current candidates into candidate table
    INSERT INTO candidate (possible_url_id, round)
    SELECT DISTINCT l.dest_url_id, round+1
    FROM link l, candidate c
    WHERE l.source_url_id = c.possible_url_id
    AND c.round = round;

    bopeep_loop(round+1, title);

ENDPROC;
```

Figure 5-18: Simple Implementation of Bo Peep

(a) Show all elements of **candidate** of the most recent round that have the sought-for title in their title field.

(b) Add the children of the newest **candidate** pages to **candidate**, with the round field set to the current value of round plus one.

(c) Let round = round + 1

Figure 5-19 shows a run of the program.

5.3.2 Technique 2: Checking with pages that referenced the old URL

Figure 5-20 shows a link containing a URL U_{bad} that is no longer valid. People who pointed to URL u_{bad} in the past are some of the most likely people to point to u_{new} now, either because they were informed of the page movement or took the trouble to find the new location themselves. Here is a heuristic based on that observation:

1. Find a set of pages P that pointed to u_{bad} at some point in the past.
2. Let P' be the elements of P that no longer point to u_{bad} anymore.
3. See if any of the pages pointed to from elements of P' are the page we are seeking.

A question is how to recognize when we've found the target page. We do this by letting the user supply a key phrase and announcing which of the linked-to pages contains that phrase in its title

? url_id('http://www.scottlan.edu/lriddle/women/BLUM.HTM');
1

SELECT value_id FROM urls WHERE url_id = 1
5

bopeep(5, "Lenore Blum");
url_id
1185

possible_url_id	vcvalue

possible_url_id	vcvalue

possible_url_id	vcvalue
63	Lenore Blum

The user manually halts the program and then requests the string associated with the matching URL...

? url(63);
http://www.scottlan.edu:80/lriddle/women/blum.htm

Figure 5-19: Transcript of Bo Peep

```
<A HREF="http://bunny.cs.uiuc.edu/funding/academicCareers.html">
ACM SIGMOD's database academic careers information</A>
```

Figure 5-20: Broken Link Appearing on "http://physio1.utmem.edu/PHYSIOLOGY/opp.html"

```
DEFPROC bopeep2(old_url_id, anchor)
    CREATE TABLE candidate (possible_url_id url_id);

    // Insert url_ids of pages that once pointed to old_url_id
    INSERT INTO candidate (possible_url_id)
    SELECT source_url_id
    FROM rlink
    WHERE dest_url_id = old_url_id;

    // Remove url_ids of pages that still point to old_url_id
    DELETE FROM candidate
    WHERE possible_url_id IN
    (SELECT source_url_id
    FROM link
    WHERE dest_url_id = old_url_id);

    // See if any of the pages pointed to by candidates have
    // the old anchor text within their title or header
    SELECT DISTINCT l.dest_url_id, v.vcvalue
    FROM candidate c, tag t, att a, valstring v, link l
    WHERE l.source_url_id = c.possible_url_id AND
    t.url_id = l.dest_url_id AND
    (t.name = 'title' OR t.name = 'h1') AND
    a.tag_id = t.tag_id AND
    a.name = 'anchor' AND
    v.value_id = a.value_id AND
    v.vcvalue LIKE strcat("%", anchor, "%");
ENDPROC;
```

Figure 5-21: Implementation of Bo Peep2

or first header. A second moved-page finder, based on this strategy, is shown in Figure 5-21. A transcript appears in Figure 5-22.

? url_id('http://bunny.cs.uiuc.edu/funding/academicCareers.html');
1253
[printed]

bopeep2(1253, 'academic careers');

dest_url_id	name	vcvalue
1362	H1	ACM SIGMOD's collection of information on academic careers and academic life
1362	TITLE	ACM SIGMOD's database academic careers information

? url(1362);
http://bunny.cs.uiuc.edu/sigmod/funding/academicCareers.html
[printed]

Figure 5-22: Transcript for Bo Peep2