Appendix C

Source Code for Home Page Finder

The home page finder is discussed in Sections 1.1.3 and 5.2.

```
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;

DEFPROC HomePageCore(fullname)
    LET fullnameExp = strcat('%', fullname, '%');

    // Get possible parents that reportedly contain name as anchor text
    INSERT INTO possibleParent (url_id, urlstring, reason)
    SELECT DISTINCT r.source_url_id, v.vcvalue, 'Anchor reportedly contains full name:
        "' + fullname + '"'
    FROM rlink r, urls u, valstring v
    WHERE r.anchor like fullnameExp
    AND u.url_id = r.source_url_id
    AND u.variant = 1
    AND v.value_id = u.value_id;

    // Find pages with *just* the full name in anchor text
    INSERT INTO candidate (url_id, score, reason)
    SELECT DISTINCT l.dest_url_id, 2, 'Anchor from "' + p.urlstring + '" is
        the full name: "' + v.vcvalue + '"'
```

```
FROM link l, possibleParent p, valstring v
WHERE l.source_url_id = p.url_id
AND l.anchor_value_id = value_id(fullname)
AND v.value_id = l.anchor_value_id;

// Find pages with the full name anywhere in the anchor text
INSERT INTO candidate (url_id, score, reason)
SELECT DISTINCT l.dest_url_id, 1, 'Anchor from "' + p.urlstring + '" includes the full
    name: "' + v.vcvalue + '"'
FROM link l, possibleParent p, valstring v
WHERE l.source_url_id = p.url_id
AND v.value_id = l.anchor_value_id
AND v.vcvalue LIKE fullnameExp;

// Increment pages with name in attribute
INSERT INTO candidate(url_id, score, reason)
SELECT t.url_id, count(*), 'The name ("' + fullname + '") appears in ' +
    CONVERT(VARCHAR(5),COUNT(*)) + ' attribute value(s) on the page'
FROM tag t, att a, valstring v
WHERE t.url_id in (select distinct url_id from candidate)
AND a.tag_id = t.tag_id
AND v.value_id = a.value_id
AND v.vcvalue LIKE fullnameExp
GROUP BY t.url_id;

// Further increment pages with name within title or header
INSERT INTO candidate(url_id, score, reason)
SELECT t.url_id, 5, 'The anchor text of a "' + t.name+ '" tag contains the full name('
    fullname + '")'

FROM tag t, att a, valstring v
WHERE t.url_id IN (SELECT DISTINCT url_id FROM candidate)
AND (t.name='title' OR t.name LIKE 'h_')
AND a.tag_id = t.tag_id
AND a.name = 'anchor'
AND v.value_id = a.value_id
AND v.vcvalue like fullnameExp;

// Find pages with URLs of the form <foo>/foo.html
INSERT INTO candidate (url_id, score, reason)
SELECT DISTINCT u.url_id, 20, 'Base of file name same as name of directory: ' + v.vcvalue
FROM urls u, parse p1, parse p2, valstring v
WHERE u.url_id IN (SELECT DISTINCT url_id FROM candidate)
AND p1.url_value_id = u.value_id
AND p2.url_value_id = u.value_id
AND v.value_id = u.value_id
AND p1.depth+1 = p2.depth
AND (p1.value = p2.value + '.html' OR p1.value = p2.value + '.htm');

// Find pages named index.html or home.html or ¡empty¿
INSERT INTO candidate (url_id, score, reason)
SELECT DISTINCT c.url_id, 10, 'File name is "' + p.value + '"'
FROM candidate c, urls u, parse p
WHERE u.url_id = c.url_id
```

```
    AND p.url_value_id = u.value_id
    AND p.depth = 1
    AND (p.value LIKE 'home.htm%' OR p.value LIKE 'index.htm%' OR
    p.value LIKE '');

    // Find pages where final directory starts with ~
    INSERT INTO candidate (url_id, score, reason)
    SELECT DISTINCT c.url_id, 10, 'Final directory name starts with tilde: "' + p.value + '"'
    FROM candidate c, urls u, parse p
    WHERE u.url_id = c.url_id
    AND p.url_value_id = u.value_id
    AND p.depth = 2
    AND p.value LIKE '~%';

    // Find pages where the penultimate directory is named
    // "home%" or "people"
    INSERT INTO candidate (url_id, score, reason)
    SELECT DISTINCT c.url_id, 10, 'Penultimate directory is: "' + p.value + '"'
    FROM candidate c, urls u, parse p
    WHERE u.url_id = c.url_id
    AND p.url_value_id = u.value_id
    AND p.depth = 3
    AND (p.value LIKE 'home%' OR p.value LIKE 'people');
ENDPROC;
```