ITR5 Information Usage

Lecture 7

Thomas Krichel

2002–03–04

Reading

Paul J. Lucas' swish++ homepage at

http://homepage.mac.com/pauljlucas/software/swish/

Source code of swish++

index

An index of a set of documents A is another documents B that contain snippets of document—or mini-documents—together with some indication on where within A these snippets can be found

Example from a book

face-to-face meetings	132–134
feedback, audience	120, 171-174

Here we use human judgement to say which snippets to use in the index.

A computer has to use simpler rules.

indexing stage: extract text

Extract ascii data out of files. if files are not straight text files, there is an extraction utility `extract` that can be run on the file.

Convert all ISO-Latin-1 characters that are not ASCII to the next ascii equivalent.

Acronyn checking

Determine whether a given word should be indexed or not using several heuristics.

First, a word is checked to see if it looks like an acronym. A word is considered an acronym only if it starts with a capital letter and is composed exclusively of capital letters, digits, and punctuation symbols, e.g., "AT&T." If a word looks like an acronym, it is OK and no further checks are done.

Second, there are several other checks that are applied.

other checks

A word is not indexed if it:

- Is less that *Word_Min_Size characters* (4) and is not an acronym.
- Contains less than *Word_Min_Vowels* (1).
- Contains more than *Word_Max_Consec_Same* (2) of the same character consecutively (not including digits).
- Contains more than *Word_Max_Consec_Consonants* (5) consecutive consonants.

- Contains more than *Word_Max_Consec_Vowels* (4) consecutive vowels.

- Contains more than *Word_Max_Consec_Puncts* (1) consecutive punctuation characters.

Note: The letter 'y' counts as a consonant both at the beginning of a word and when it has a vowel in front of it; otherwise (when it follows a consonant), it is treated as a vowel.

stop words

A word is not indexed if it belongs to a built-in list of stop word:

"about", "above", "according", "across", . . . "yes","yet", "you", "you'll", "your", "you're", "yours", "yourself", "yourselves", "zero".

At the start of indexing the built-in list is uesd. If and index is there, then the list of its stop word is used. By default, the indexer discards all words as stop words that occur in *Word_Percent* (100) of all files. If that value takes 101, no words are discarded, except those in the built-in list.

Problem with incremental indexing...

html,xhtml indexing mode

- Character and numeric (decimal and hexadecimal) entity references are converted to their ASCII character equivalents before further examination and indexing. For example, the word `r&eacute;sum&#233;` becomes "resume" before indexing.

- If a matched set of `<TITLE>` ... `<TITLE>` tags is found within the first TitleLines lines of the file (default is 12), then the text between the tags is stored in the generated index file as the file's title rather than the file's name.

- If an HTML or XHTML element contains a CLASS attribute whose value is among the set of class names specified as those not to index then all the text up to the tag that ends the element will not be indexed.

html,xhtml indexing mode

- If an HTML or XHTML element contains a TITLE attribute, then the words specified as the value of the TITLE attribute are indexed.
- If an AREA, IMG, or INPUT element contains an ALT attribute, then the words specified as the value of the ALT attribute are indexed.
- If a META element contains both a NAME and CONTENT attribute, then the words specified as the value of the CONTENT attribute are indexed associated with the meta name specified as the value of the NAME attribute.

html,xhtml indexing mode

- If a TABLE element contains a SUMMARY attribute, then the words specified as the value of the SUMMARY attribute are indexed.
- If an OBJECT element contains a STANDBY attribute, then the words specified as the value of the STANDBY attribute are indexed.
- All other HTML or XHTML tags and comments (anything between < and > characters) are discarded.