December 1998

A Knowledge-Based System for Improving WWW Search Results

Radhika Santhanam
University of Cincinnati

Anthony Scime
Wilson College

Joyce Elam
Florida International University

A Knowledge-Based System for Improving WWW Search Results

Anthony Scime'
Business and Economics Department
Wilson College

Abstract

A Web search engine's returned hits are not always ordered to meet the needs of the user. The search results can be improved using heuristics about the structure of Web sites and domain knowledge provided by the user. This paper presents a method for the refinement and reordering of search engine results.

Introduction

Web search engines determine a site's quality in response to a search request by matching keywords in the request to keywords representing the site. The resulting sites, hits, are given a score and ranked according to the match of the keywords [KOW97]. The keywords in the query represent the information needs of the user. The resulting score is the value of the site to the needs of the user [TUR96].

Search engines generally list hits in score order. Scoring can be done based on the number of keywords found, keyword frequency, or keyword location [KOW97]. Yet, the ranking systems continue to provide large numbers of sites, and sites not relevant to the user's search request.

This paper proposes a repertory grid rating method to improve the ranking of search results. This method is based upon two constructs. A syntactic evaluation scores a site's classification as a value independent of the current search. A semantic score measures the site's fit to a search hierarchy. The repertory grid combines the two scores to determine a site's overall score. For example, a search to find office equipment companies in Florida can be viewed as a specialization hierarchy that acts as a description of the user's knowledge in the domain. The root node of this hierarchy is "Florida." This search is interested in Florida businesses, providing the intermediate node "business." Specifically office equipment suppliers are sought. This provides the keyword for the search – "Office Equipment." Figure 1 provides the complete hierarchy.

Florida ——— Business ——— Office Equipment

Figure 1. Florida–Business –Office Equipment Hierarchy

Syntactic Evaluation of a Site

The syntactic evaluation is a determination of a site's type. Sites on the Web can be classified on a continuum of possible types – direct hit, page hit, directory hit. These site classifications are general classes to which the sites can be bound.

Pages that provide significant information about a site's content, enough perhaps to satisfy a user's search are a direct hit. Often home pages are direct hits. Sites that are subordinate to a home page belong to the general class of page hit. Typically, these sites contain information about part of the parent's subject. Directory hits are sites that offer links to other sites. The directories generally do not provide information about the subject of the site. Their primary purpose is to lead the user to information containing sites provided by the directory as a service. Directories may be organized by subject, geography, or any other method.

The classifications are not absolute. Many sites may have characteristics of more than one type. For example, a site may belong to a collection of sites, contain a significant amount of information about its topic, and provide links to other sites with even more information. This mixing of classifications creates the continuum on which sites are classified.

To visit each site returned by a search engine is an overwhelming task. Evaluation of a site's type is determined by a syntactic review of the structure of the site's URL and title. The URL generally provides an indication of the site's relative importance in the hierarchy of the site's parent. A URL is composed of the Web server address and the directory structure to the site. The longer the URL, the further away the site is from the server, the root indicating more specialization of content. The title of a site and its relationship to the URL provides a further indication of type. A review of sites, identified similarities that were developed into the heuristics shown in Figure 2.

Direct hit rules

1. A site tends to be a direct hit when there is a word in the title that is a string in the URL before the name domain (i.e., .com, .edu, .gov, .mil, .us, .ca, .au, etc.). This rule does not apply if the word is "www".
2. A site tends to be a direct hit when the domain name is the last group of characters in the URL.
3. A site tends to be a direct hit when "home" is a word in the title or a string in the URL.

Page hit rules

1. A site tends to be a page hit when there is a word in the title that is a string in the URL after the domain name.
2. A site tends to be a page hit when the URL does not end in ".htm" or ".html".
3. A site tends to be a page hit when a numeric digit is between 5 and 12 places from the end of the URL.
4. A site tends to be a page hit when the string "pg" occurs in the URL.

Directory hit rule

A site tends to be a directory hit when one of the words "directory," "add," "ads," "classified," "sponsors," "members," "mall," "index," or "menu" appears as a string in the URL, as a word in the title, or as a word in the summary.

Figure 2. Syntactic Rules

By assigning numeric values to the site types, each rule can provide a syntactic score. A "direct hit" scores 5, page hit 3, and directory hit 2. Sites that cannot be classified are assigned a score of 1. Multiple classifications by different rules are averaged to place the sites along the site syntactic continuum.

For example, in the search for office equipment providers in Florida a site for the Southern Office Equipment Company was found. The company's home page is titled "Southern Office Equipment Home Page" with a URL of http://www.soffice.com/sofhomel.htm. Using the syntactic rules (Figure 2) this site is a "direct hit" by direct hit rule 1, a "direct hit" by direct hit rule 3, and a "page hit" by page hit rule 1. Overall this site scores 4.333 on the syntactic continuum, by averaging the individual scores (5 + 5 + 3 = 4.333, Figure 3).

Semantic Evaluation of Sites

The semantic evaluation is a measure of the number of nodes along the hierarchy, which match a site description. The more nodes in common with the site, the greater the value. Dimensional analysis is used to convert the number of nodes into an ordinal ranking. This allows an accurate comparison of site semantic values [MAR95].

$$r_{alternative} = 10 - \{[(v_{alternative} - v_{max}) / (v_{min} - v_{max})] * 10\}$$

where:

$r_{alternative}$ is the score of the site,
$v_{alternative}$ is the number of hierarchy nodes in common with the site,
v_{min} is zero, the minimum number of hierarchy nodes possible to match, and
v_{max} is the number of nodes in the hierarchy.

The semantic score for a site will fall along the scale 0 to 10. A semantic evaluation was conducted using the hierarchy – Florida – Business – Office Equipment. The Southern Office Equipment Home Page summary contains two of the nodes in the taxonomy - "Office Equipment" and "Florida." The summary is as follows:

For the high-graphics version of our site, click here. For our text-only version, click here. Welcome to Southern Office Equipment. 4424 North Lois Avenue * Tampa, Florida 33614 (800) 555-7637 * Fax (813) 555-7517 You can also send us E-mail at:

Using the dimensional analysis: $r_{alternative} = 10 - \{[(2 - 3) / (0 - 3)] * 10\} = 6.6667$,
and Southern Office Equipment Home Page falls on the semantic construct line as shown in Figure 4.

Repertory Grid

To determine a site's overall relevance to a user's search request, the product of the syntactic and semantic scores is computed using a repertory grid [SHA87]. The repertory grid score for the Southern Office Equipment Home Page, 28.8889, is the product of the syntactic score, 4.3333, and the semantic score, 6.6667 (Figure 5). Performing this analysis on all sites returned by a search creates a repertory grid with rankings considering the hits quality and content.

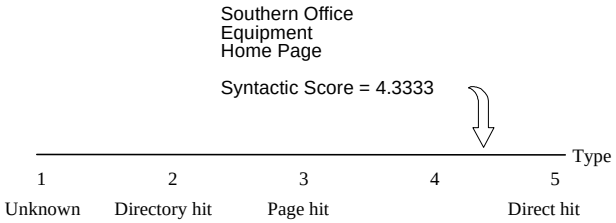

Figure 3. Syntactic Site Continuum

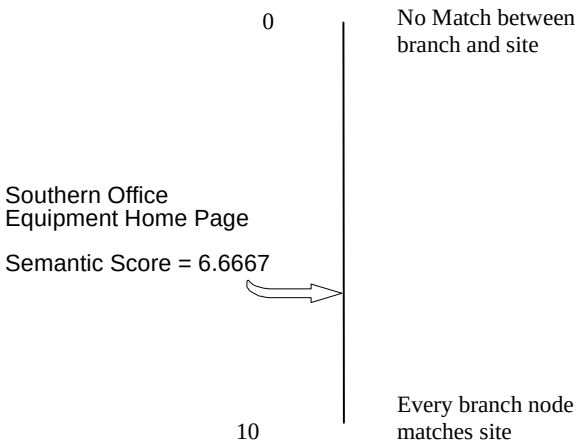

Figure 4. Semantic Construct Line

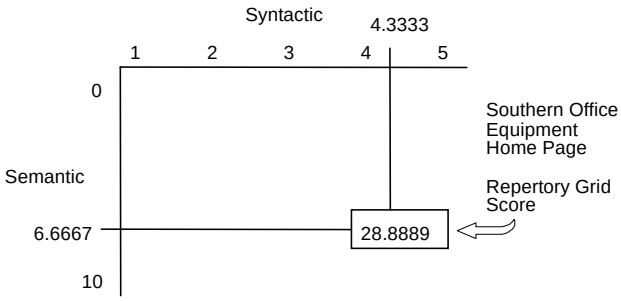

Figure 5. Repertory Grid

Results

In the Florida office equipment search, the syntactic and semantic scores combine to create a ranking approach that reorders the original search engine results (Table 1). The direct-hit and page-hit combination for "Southern Office Equipment" gave the highest score, up from sixth place originally. The site "gotwh (sic) energy efficient office equipment," which is not in Florida, does not sell office equipment, and is a directory of other sites, correctly received the lowest score (down from first place). The top two sites are Florida office equipment businesses (originally sixth and seventh), followed by a directory of office equipment businesses in Florida (down from second). The fourth, and fifth sites (up from eight and ninth) are office equipment businesses outside Florida. The seventh site (down from third) is the first one not related to selling office equipment.

Two scoring methods are combined to do the site ranking. The first, syntactic, uses heuristics to determine the quality of the site's information. The second, semantic, considers the user's knowledge of the domain from which the search keyword is derived. This permits an improvement over the search engine ranking based solely on the keyword's relationship to the site.

References

[SHA87] Shaw, M. L. G., and Gaines, B. R.; "KITTEN: Knowledge Initiation and Transfer Tools for Experts and Novices;" *International Journal of Man-Machine Studies*; September 1987; pp. 127-134.

[TUR96] Turtle, Howard R., and Croft, W. Bruce; "Uncertainty in Information Retrieval Systems;" in *Uncertainty Management in Information Systems From Needs to Solutions*, ed. by Motro, A. and Smets, P.; Kluwer Academic Publishers; Boston; 1996; pp. 436-452.

[MAR95] Martin, M.; *Analysis and Design of Business Information Systems*; 2nd Ed.; Prentice-Hall; Englewood Cliffs; 1995; pp. 290-291.

[KOW97] Kowalski, Gerald; *Information Retrieval Systems: Theory and Implementation.* Kluwer Academic Publishers; Boston; 1997; pp. 149-177.

Table 1. Florida–Business–Office Equipment Results

Semantic Score	Syntactic Score	Repertory Grid Score	Title	URL	Original Rank
6.6667	4.3333	28.8889	southern office equipment home page	http://www.soffice.com/sofhomel.htm	6
6.6667	4.3333	28.8889	southern office equipment home page	http://www.soffice.com/sofhomem.htm	7
10	2.5	25	directory of south florida office equipment exporters	http://www.exportmiami.com/office/office.htm	2
6.6667	3	20	office equipment stationary	http://www.f3.com/fairs/13/2/june.html	8
6.6667	2.5	16.6667	office equipment stationary “&” supplies importing firms	http://www.export-leads.com/cd02e.html	9
3.3333	4	13.3333	column office equipment locations throughout chicagoland	http://www.columinc.com/location.htm	3
3.3333	4	13.3333	project green machine cost effective energy savings with computers and office	http://pacwww.chch.cri.nz/ema/co_prof/paper96/135/ba.htm	5
3.3333	3.6667	12.2222	state of minnesota computer and office equipment employment	http://peter.itsc.state.md.us:81/aces/mn335703.htm	10
3.3333	3.3333	11.1111	findit office equipment index	http://www.findit.co.uk/office.htm	4
3.3333	2.5	8.3333	gotwh energy efficient office equipment	http://crest.org/environment/gotwh/general/off-equip/index.html	1