25만 서버, 3천만 웹페이지 인덱스 보유

꼭 필요한 정보 '알타비스타'로 …

디지탈社의 알타비스타는 현재 인터넷상에서 가장 빠른 검색 서비스를 제공하는 검색엔진이라 할 수 있다. 25만여 서버와 3천만개 이상의 웹페이지 인덱스를 보유하고 있는 동시에 1만 4천개 이상의 뉴스그룹에서 300만개 이상의 기사를 검색할 수 있다. 매일 알타비스타를 사용하는 건수도 1,200만건에 이르고 있는 알타비스타(URL : www.altavista.digital.com)에 대해 살펴봤다. 〈편집자〉

알타비스타는 중대형 컴퓨터 공급업체인 미국의 디지탈사에서 제공하는 검색엔진으로, 현재 가장 많은 웹 인덱스를 가지고 있다. 95년 여름 캘리포니아주의 팔로알토에 있는 디지탈의 리서치랩에서 연구를 시작하여 인덱스 구축작업을 한 뒤 약 2개월간의 내부 시험을 거쳐 95년 12월 중순부터 일반에게 서비스를 제공하기 시작했다.

서비스를 시작한 후 몇 달동안 많은 데이터베이스를 구축하여 지속적인 확장을 꾀한 알타비스타는 현재 인터넷상의 25만여 서버와 대략 3천만개 이상의 웹페이지에 대한 인덱스를 가지고 있다. 또한 1만 4천개 이상의 뉴스그룹에서 300만개의 기사를 검색할 수 있기 때문에 매일 접속하여 검색하는 사용자 수가 약 1,200만건 이상에 이르고 있다.

알타비스타는 이처럼 많은 양의 데이터베이스와 사용자 접속에도 불구하고 아주 빠른 속도로 검색이 이뤄지고 결과의 전송 또한 매우 빠르기 때문에 인터넷 사용자들의 많은 사랑을 받고 있다. 실제로 거의 모든 질의가 1초 이내에 실행되며 전송속도도 최상의 속도로 이루어지고 있어 현재로서는 인터넷에서 가장 빠른 최고의 검색엔진이라 할 수 있다. 알타비스타에서 사용되고 있는 하드웨어와 소프트웨어는 다음과 같으며, 많은 장비를 갖추고 있는 알타비스타는 가장 빠르고 많은 정보를 제공하는 서비스로 자리잡았다.

· 외부 트래픽 관리 : 256MB 메모리와 4GB의 하드디스크를 가지고 있는 알파스테이션 500
· 웹 인덱서 : 6GB의 메모리와 210GB의 레이드 디스크를 가진 알파서버 8400 5/300으로 대부분의 검색요청을 1초 이내에 수행
· 스쿠터 : 1GB의 메모리와 30GB의 레이드 디스크를 가진 DEC 3000/900웍스테이션으로 전세계의 웹을 검색하여 웹 인덱서에 자료를 제출
· 뉴스 인덱서 : 196MB의 메모리 13GB의 디스크를 가진 알파스테이션 250 4/266
· 뉴스 서버 : 160MB 메모리 24GB의 레이드 디스크를 가진 알파스테이션 400 4/233

알타비스타 검색 방법

· Search : 검색할 문서의 종류를 결정하는 것으로, 검색대상을 웹사이트로 할 것인지 뉴스 그룹으로 할 것인지 결정한다.
  - the Web - 검색 대상을 웹사이트로 지정
  - Usenet - 검색 대상을 뉴스 그룹으로 지정
· Display the Results : 사용자에게 되돌려지는 검색 결과의 형태를 어떻게 표시할 것인지 결정하는 것으로, 결과로 돌아오는 웹사이트의 URL에 꼬리표처럼 붙어 있는 설명을 어떻게 표시할 것인지를 결정한다.
  - in Standard Form - 표준 형태
  - in Detail Form - 상세한 설명
  - in Compact Form - 간단한 설명
  - Sumit 버튼 - 구성 완료된 질의를 제출하는 버튼

- Simple Search - 일반적 검색
- Advance Search - 이진 오퍼레이터를 사용하는 검색

이것은 검색방법을 결정한다기보다 질의를 제출하는 방법에 대해 결정하는 것이다. 즉, 심플 서치와 어드밴스드 서치는 동일한 검색엔진에 대해서 인터페이스만 달리하는 것이다. 따라서 질의에 대한 결과 순위를 결정할 때 어떤 식으로 결정할 것인지에 영향을 미친다.

심플 서치와 어드밴스드 서치는 질의를 제출할 때 각각 다른 종류의 오퍼레이터들을 사용한다. 심플 서치에서는 따옴표와 +, - 등의 기호를 사용하고 어드밴스드 서치에서는 이진 오퍼레이터인 AND, OR, NEAR, NOT, &, |, ~, ! 등 주로 프로그램 언어에서 사용하는 오퍼레이터를 그대로 사용한다.

간단한 질의를 제출할 때에는 심플 서치가 편리하지만 실제로 아주 복잡한 질의를 제출할 때는 이진 오퍼레이이터를 사용하는 어드밴스드 서치가 더 편리하다. 그러나 실질적으로 모든 심플 서치는 어드밴스드 서치 형태로 변환되어 서버에 제출된다.

검색 결과의 표시

키워드를 제출하면 알타비스타는 그 단어가 들어있는 문서를 찾아낸다. 알타비스타는 일단 제출된 단어가 들어있는 모든 문서를 검색한 다음 단어가 나타나는 횟수와 함께 각 문서들을 특정한 방법으로 순위을 매겨 그 순위에 따라서 화면에 차례대로 나타나게 한다.

즉, 횟수는 알타비스타가 가지고 있는 데이터베이스에 제출된 단어가 나타나는 총 횟수를 말한다. 그리고 점수를 매기는 알고리즘은 다음과 같다. 점수의 비중은 1번이 가장 높고 그 다음 2번, 다음 3번 순으로 나타난다.

· 질의 단어 또는 문장이 나오는 순서 : 질의로 제출된 단어 또는 문장이 앞에 나올수록 높은 점수를 얻는다. 예를 들어, 질의 단어가 웹페이지의 타이틀이나 뉴스 그룹의 헤더에 있는 문서가 높은 점수를 받는다.
· 질의 단어들이 서로 가까이 붙어 있는 순서 : 문서에서 제출된 단어나 문장이 서로 가까이 붙어 있을수록 높은 점수를 받는다.
· 질의 단어들이 나타나는 횟수 : 횟수가 높을수록 높은 점수를 얻는다.

· 질의 제출과 조건 설정
- 먼저 검색할 문서 대상을 정한다. 즉, the Web 또는 Usenet 중 한가지를 선택한다.
- 심플 서치와 어드밴스드 서치 중에서 원하는 것을 선택한다.
- 마지막으로 결과 표시 방법을 선택한다.

질의 제출 방식

심플 서치 : 일반적인 검색보다 일상 언어에 가까운 오퍼레이터들을 사용하여 질의를 제출한다. 심플 서치에 사용되는 오퍼레이터들은 " "와 +, -, * 등이다.

겹따옴표 기호는 여러개의 단어를 한개의 단어처럼 취급하도록 만든다. 이는 겹따옴표 사이에 들어있는 모든 단어들을 집합적으로 하나의 단어처럼 취급한다. 키워드를 제출할 때 아무런 지시없이 단어들을 쭉 나열해서 제출하면 알타비스타는 그 단어들이 들어있는 모든 문서를 찾는다. 따라서 아무런 지시없이 두 개의 단어를 키워드로 제출하면 그 단어들이 각각 들어있는 문서들을 모두 찾게 되어 검색결과로 돌아오는 인덱스의 양이 엄청나게 많아진다.

물론 그것들 중에서 두 단어가 서로 붙어있는 것에 높은 점수를 줘서 결과를 돌려보내지만 두 단어가 서로 연결되어 있는 것이라는 사실을 명확하게 하려면 겹따옴표 " "로 묶어서 제출한다. 그러면 알타비스타는 따옴표 안에 있는 것은 마치 모두 한 단어인 것처럼 취급하여 그 단어들이 서로 나란히 붙어서 들어있는 문서들의 인덱스만을 돌려보낸다.

겹따옴표의 사용은 여러개의 단어를 키워드로 제출할 때 그 중 몇 개의 단어만이 서로 붙어있는 경우에 필수적이다. 예를 들어, hotel "ocean front" - 해변에 있는 호텔을 검색할 때 ocean front는 두 개의 단어지만 숙어로서 하나의 단어처럼 인식되어야 할 필요가 있다. 반면 hotel은 이 질의에서 ocean front와 꼭 붙어 있어야만 하는 것은 아니다.

또한 단어가 서로 붙어있는 문서들의 URL이 높은 점수를 얻기는 하지만, 실질적으로 질의를 제출해 보면 겹따옴표로 묶은 것과 다른 결과가 나타나는 경우가 많다. 일반적으로 겹따옴표로 묶었을 때 자신이 원하는 것을 찾아낼 가능성이 훨씬 크다.

예를 들어, 키워드로 서치 엔진을 제출한 경우와 "서치

엔진"을 제출한 경우를 서로 비교해 보면 그 차이를 이해할 수 있을 것이다.

+는 일명 requirment라고도 한다. +기호 다음에 있는 단어가 들어가 있는 문서만 돌려준다. -는 prohibit 즉 이 기호 다음에 있는 단어가 들어있지 않은 문서만 결과로 돌려준다. 예를 들어, korea +south -north와 같이 하여 질의를 제출하면 korea란 단어가 들어있는 문서중에 south가 들어 있으며 north가 들어있지 않은 문서만 결과로 돌려준다.

*는 와일드 카드 문자. 예를 들어, korea*를 키워드로 제출하면 korea, korean, koreans, koreana 등을 모두 찾아준다. 이 와일드 카드의 사용은 실제로 미국과 영국에서 표기를 달리하는 단어를 검색할 때 사용하면 특히 편리하다. 예를 들어, alumi*m을 입력하면 미국식 표기인 aluminum과 영국식 표기인 aluminium을 모두 검색할 수 있다. 또는 theat*를 검색어로 사용하여 theater와 theatre가 들어 있는 문서를 모두 검색할 수 있다.

그러나 이 와일드 카드 문자를 사용할 때는 주의할 필요가 있다. 너무 많은 결과가 돌아오지 않도록 해야 한다. 알타비스타에서는 결과 인덱스가 너무 많으면 그 키워드는 무시하게 된다.

대문자와 소문자 : 소문자만을 사용하면 대문자와 소문자를 구별하지 않는다. 예를 들어, korea를 키워드로 제출하면 korea, Korea, KOREA가 들어있는 문서들을 모두 찾아준다. 키워드에 대문자를 사용하면 정확하게 대문자와 소문자를 구별하여 검색한다. 그렇지만 korEA라고 입력을 하면 정확하게 korEA란 단어가 들어 있는 문서만을 검색한다. 그러므로 특별한 경우를 제외하고 일반적으로 검색할 때에는 소문자만을 사용하는 것이 편리하고 안전하다.

기호에 대해서

알타비스타에서 여러가지 기호들을 포함시켜 검색할 때에는 세심한 주의가 필요하다. 이것은 알타비스타의 데이터베이스에서는 오직 단어들만이 의미가 있고, 단어를 구분하는 기호나 알파벳이 아닌 기호들로서 $, %, /, ~, # 등은 모두 단어를 구분하는 것을 의미할 뿐이다. 따라서 다음은 모두 각각 6단어로 된 구문이며 동일하게 취급된다.

President of the U.S.A
President-of-the-U-S-A
President/of/the/U/S/A
President. of. the. U-S-A
President of the U S A

또한 어포스토로피도 단어를 구분하는 기호로 인식된다. 예를 들어, Kim's Restaurant라고 입력하면 이것은 실제로 3단어로 인식된다. 또한 AT&T, 3.141592, don't, digital.com.x-y 등은 모두 2단어로 인식된다. 주의할 것은 &, |, !, ~ 등은 어드밴스드 서치에서 중요한 의미를 갖고 있음을 꼭 기억해야 한다.

어드밴스드 서치는 연산자와 표현 인덱스, 즉 이진 오퍼레이터들을 사용하여 질의를 구성한다. 알타비스타 어드밴스드 서치에서 사용하는 오퍼레이터들은 다음과 같다.

- AND, & : AND로 연결되는 단어가 모두 들어있는 문서 검색. 예를 들면, rafting AND australia의 경우는 오스트리아와 래프팅이 모두 들어있는 문서를 검색할 때 사용하는 표기다.
- OR, | : OR로 연결되는 단어 가운데 적어도 하나가 들어있는 문서 검색. 예를 들면, theater or theatre의 경우, theater가 들어있거나 theatre가 들어있는 문서 또는 둘 다 들어있는 문서를 모두 검색할 때 사용한다.
- NEAR, ~ : NEAR로 연결되는 단어들이 서로 10단어 이내에 위치한 문서 검색.
- NOT, ! : NOT 뒤에 나오는 단어가 없는 문서만을 검색.

이 오퍼레이터들은 소문자로 사용해도 된다. 즉, and, or, near, not 등과 같이 사용할 수 있다. 만약 오퍼레이터와 같은 단어가 들어있는 문서를 검색하려면 겹따옴표 "" 로 묶어야 한다. 예를 들어, AT&T를 검색하려면 "AT&T""로 질의를 제출해야 한다.

또한 괄호를 사용하면 혼동을 방지할 수 있다. 예를 들어, rock and not "dance music" 보다는 rock and(not "dance music")이 같은 의미면서 보다 뜻을 명확히 할 수 있다. 물론 서버가 혼동하는 일은 없지만 질의를 제출하는 사용자 입장에서 복잡한 질의를 구성하다 보면 괄호를 사용하는 것이 자신의 의도를 명확하게 하여 오류를 방지할

수 있다. 한편, 단어와 구문, 대문자 사용 와일드 카드 문자 사용 등은 심플 서치와 동일하다.

어드밴스드 서치에서는 중요하게 다루어져야 할 것이 3개 이상의 단어들이 오퍼레이터를 이용하여 키워드로 구성될 때 단어들의 결합 우선 순위이다. 어드밴스드 서치에서는 이진 오퍼레이터들을 사용하므로 사칙연산에서 연산의 우선 순위와 마찬가지로 단어들의 결합 우선 순위가 결정된다. 또한 괄호로 묶인 단어들이 우선적으로 결합된다. 그러므로 괄호를 적절하게 잘 사용하면 강력한 질의를 구성할 수 있다.

( )의 사용에 대해서

( )의 사용은 3개 이상의 단어들이 오퍼레이터를 이용하여 키워드로 구성될 때 단어들의 결합 우선 순위를 지정하는데 사용된다. 중고등학교 때 수학에서 사칙연산의 우선 순위를 지정하는 것과 같이 오퍼레이터의 효력과 우선 순위를 지정하는 것이다. 또한 괄호로 묶인 단어들이 우선적으로 결합된다는 사실도 기억해 두자.

일반적으로 키워드 검색에서 질의를 제출하면 얻어지는 결과물에는 원하는 정보들과 함께 원하지 않는 정보들도 섞여있는 경우가 거의 대부분이다. 때문에 열거되어 있는 경우가 거의 대부분이다. 그러면 열거되어 있는 URL을 하나씩 방문해 봐야 그것이 자신이 원하는 것인지 아닌지 알 수 있다.

그러므로 될 수 있는대로 검색 결과를 축소하는 것이 유리하다고 볼 수 있는데 여러개의 단어를 키워드로 제출할 때 ( )를 적절하게 사용하면 질의를 아주 정밀하게 만들어 시간을 많이 절약할 수 있게 해준다. 조금 복잡하지만 다음의 예들을 각각 비교하면서 ( )의 사용법을 익혀보자.

apple or orange and lemmon
apple or (orange and lemmon)
(apple or orange) and lemmon

첫번째와 두번째는 동일한 질의로서 사과를 포함하는 문서들과 더불어 오렌지와 레몬이 동시에 들어있는 문서를 결과로 돌려준다. 세번째 질의는 사과와 레몬이 함께 들어 있는 문서와 오렌지와 레몬이 함께 들어있는 문서를 결과로 돌려준다.

not, apple and orange
(not apple) and orange
not (apple and orange)

여기서도 첫번째와 두번째는 같은 질의로서 오렌지는 들어있지만 사과는 들어있지 않은 문서를 결과로 돌려준다. 세번째 질의는 사과와 오렌지가 둘다 들어있지 않은 문서들을 결과로 돌려준다.

apple near orange and lemmon
(apple near orange) and lemmon
(apple near orange) and (apple near lemmon)

첫번째와 두번째는 동일한 질의로 오렌지에서 10개의 단어 이내에 사과가 들어있으며, 동시에 레몬이 들어있는 문서를 결과로 돌려준다. 사과라는 단어와 오렌지라는 단어가 서로 10단어 이내에 있으며 동시에 사과와 레몬이 10단어 이내에 들어있는 문서를 결과로 얻고 싶다면 세번째 질의를 사용하면 된다.

not apple near orange
not (apple near orange)
apple and not (apple near orange)

첫번째 두번째 질의는 동일. 사과와 오렌지가 서로 가까이 붙어있는 문서를 제외시킨다. 세번째 질의는 사과가 들어있기는 하되 오렌지와 가가이 붙어있지 않은 문서만을 검색하도록 한다.

알타비스타의 이런 질의 제출에 익숙하지 않은 사람들은 처음에 질의 구성에 시간이 많이 걸릴 수도 있겠으나 몇번 해보면 금방 익숙해져서 강력한 질의를 제출할 수 있게 될 것이다.

따라서 익숙해진 질의 제출을 통해 수많은 URL을 검색함으로써 사용자가 원하는 자료를 보다 수월하게 습득할 수 있도록 알타비스타는 사용자 개개인에게 안내자가 될 것이다. DC