'97 International Conference Multimedia Databases on Internet
October 10, 1997, Seoul Korea

Web 상에서의 정보검색기법에 관한 고찰

김준오 (㈜유진데이터 기술연구소)

요약

현재 기업 뿐 아니라 개인에게도 정보의 중요성이 증대되고 있는 실정이다. 특히, 정보기술의 활용과 빠른 컴퓨팅 환경의 변화 및 인터넷이라는 네트웍을 기반으로 다양한 정보를 접할 수 있게 됨으로써 필요한 정보를 보다 빠르게 검색해야 하는 필요성이 생기게 되었다.

지금까지는 다양하고 방대한 데이터들을 정보의 형태로 가공하여 DB 를 중심으로 유용한 정보를 제공하였지만, 이제는 효과적인 정보의 활용을 위해 정보검색의 중요성이 증대되고 있다. 즉, 흩어져 있는 데이터를 정보의 형태로 변환시키는 것보다 그 정보를 효율적이고 빠르게 검색하여 한차원 높은 'Knowledge' 로써의 역활을 하느냐가 중요하게 된것이다.

본 논문에서는 각 정보들의 검색을 위해 사용되고 있는 기존의 검색기법인 SQL-Based 검색, Full text 검색과 새로 소개되고 있는 Parametric 검색에 대해서 고찰하여보고 다양한 정보의 유형에 대해 효과적인 검색을 위한 방안을 제시하고자 한다.

1. 서론

최근 컴퓨팅환경은 빠른 변화를 맞이하고 있으며, 최신의 정보기술들이 소개되고 있다. 이들 컴퓨팅환경의 변화와 정보기술들의 내용을 살펴보면 다음과 같다.

- Client/Server 환경 ▶ Web-based 환경
- Transaction Processing ▶ Information Retrieval & Information Analysis
- Distributed Computing 환경
- System Integration
- Java
- CORBA
- Object-Oriented Technology
- Internet/Intranet/Extranet

위와 같이 컴퓨팅 환경은 기존의 Client/Server 환경에서 Intranet이라는 Web을 기반으로 하는 정보시스템이 등장하고 있으며, 데이터 처리 중심의 환경에서 정보의 검색 중심으로 변화하고 있다. 특히, 이들 정보들을 물리적인 위치에 상관없이 사용할 수 있도록 분산환경을 지원하고자 하고 있으며, 개별적인 시스템 환경이 아닌 여러 정보시스템 또는 여러 정보원들과의 유기적인 연계를 위해 시스템의 통합화가 추진되고 있다. 또한, 이러한 시스템을 네트웍 환경에서 효율적으로 구현하고자 Java와 CORBA등이 사용되고 있으며, 여러 분야에 걸쳐 객체지향 개념들이 활용되고 있다.

이와 같은 컴퓨팅 환경의 변화와 정보기술들의 활용은 기업내의 폭발적인 정보의 증가를 초래하였다. 따라서, 다양한 데이터에서 정보를 생성하고 효율적으로 관리하는 지금까지의 환경뿐 아니라 생성된 정보를 얼마나 효율적으로 사용할 것인가 또한 중요시 되고 있다.

이와 같은 정보의 Digital 환경에서는 Internet을 기반으로 생성된 정보에 대해 방대한 정보의 공유가 가능하게 됨으로써 사용자가 원하는 정보의 효율적인 검색할 수 있는 정보검색기술이 부각되고 있다. 이제 정보는 흩어져 있는 데이터들을 가공된 형태로 제공하는 것 뿐만 아니라 실제로 사용자가 필요한 정보를 쉽고 빠르게 검색하여 활용할 수 있는 Knowledge로써의 역할이 기대되고 있다.

2. Web 상에서의 정보검색

웹을 통하여 방대한 양의 정보들이 전 세계적으로 공유되고 있으며, 이제는 문서의 형태로 웹사이트에 제공되는 정보뿐 아니라 내부의 DB로 관리되고 있던 정보시스템도 웹서비스로 발전하고 있는 추세이다.

웹이라는 인터넷 서비스를 활용함으로써 정보의 요구가 더욱 다양해지고, 기존의 텍스트 중심의 정보에서 벗어나 다양한 유형의 데이터 즉, 텍스트, 오디오, 동영상, 비디오 정보 등의 멀티미디어 데이터와 학술, 문헌, 여행, 상품, 구매 등의 다양한 정보들이 망라되어 제공되고 있다.

이제는 단순히 검색엔진을 통하여 웹사이트를 검색하는 수준이 아닌 웹상의 정보 및 웹과 연동되어 제공되어지는 모든 공간에 존재하는 다양한 형태의 정보들을 효율적으로 검색할 수 있어야 할 것이다.

정보검색을 위해 다음과 같은 분류 기준을 고려하고 있다.

- 검색방법에 따른 분류
- 검색목적에 따른 분류
- 검색대상에 따른 분류

가. 검색방법에 따른 분류

웹상에서 일반적으로 사용되는 검색엔진을 기준으로 키워드 검색과 주제어 검색으로 분류하고 있다. 키워드 검색은 사용자가 주제어가 되는 키워드를 입력함으로써 해당 키워드가 들어있는 모든 정보를 검색하는 것이고, 주제어 검색은 미리 검색엔진에서 주제별로 분류해서 설정해 놓은 정보를 디렉토리 개념으로 검색해 가는 방식이다. 현재는 웹사이트의 폭발적인 증가에 의해 DB의 규모가 커짐에 따라 키워드 검색이 많이 사용되고 있다.

이제는 단순히 웹사이트외에 다양한 정보가 제공되고 있으므로 보다 많은 검색기법들이 고려되어야 할 것이다.

나. 검색목적에 따른 분류

정보검색을 하는 목적에 따라 단순검색과 전략검색으로 분류하고 있다. 단순검색은 찾고자 하는 정확한 하나의 정보를 검색하는 것으로써 정보사냥대회처럼 정확한 검색을 목적으로 하는 것이고, 전략검색은 다양한 정보원에 대한 포괄적인 정보검색을 통하여 분석적 검색을 하는 것으로써 기술동향, 특허, 비교분석적 정보를 찾는 것이다.

다. 검색대상에 따른 분류

검색대상에 따라 웹상의 존재하는 정보와 전문DB로 구축되어 있는 정보로 분류할 수 있다. 특히, 전문DB의 경우 웹상에서 그 정보가 있는 곳의 위치를 파악하기 전에는 검색이 불가능하다. 하지만, 웹과 연동되는 활용가능한 정보원으로써 전문DB도 웹상에서 검색이 가능하게 되었고, 앞으로 그 수는 증가하고 있는 추세이므로, 웹상에서도 중요한 검색대상이 될 것이다.

3. SQL-Based Search

가. SQL-Based Search 의 개요

일반적으로 DBMS로 구축된 정보를 표준 질의문인 SQL을 통하여 검색하는 것을 말한다. 기업내의 정보관리시스템 또는 상용 DB 서비스를 위해 구축된 DB정보를 검색하는데 사용되고 있으며, 지금까지 정보검색을 위해 가장 많이 사용되는 검색방법이라고 할 수 있다.

나. SQL-Based Search 의 특징

기업내의 정보에 대해 transaction을 위해 구축되거나 상용서비스를 위해 구축되어 있는 DB를 검색하는 것으로 일반적으로 정형화된 정보를 대상한다. 특히, 정보의 분석 또는 검색을 위한 DB를 제공하는 것이 아니라 처리 또는

관리를 위한 것이고 또한 정보의 구조가 2차원적인 동일한 구조에 의해 제공되므로 제한적이라 할 수 있다.

4. Full Text Search

가. Full Text Search 의 개요

Full text search는 입력된 질의문을 분석하여 이를 색인어 어휘로 변환한 후 색인어가 인덱싱되어 있는 DB를 탐색하여 해당 문서 또는 해당 웹사이트를 검색하게 된다.

일반적으로 Full text search는 문헌정보 검색에 많이 쓰이고 있으며, 웹상에서의 여러 검색엔진등이 웹사이트내의 HTML문서 정보를 찾아내기 위하여 사용하고 있다. 문서의 검색의 경우는 찾고자 하는 단어를 입력함으로써, 모든 문서를 수동으로 찾아볼 필요없이 해당 단어가 포함되어 있는 것들을 찾아주게 된다. 특히, 웹상에서의 몇몇 검색엔진들은 검색 Robot을 이용함으로써 웹상에 등록되어 있는 웹사이트를 검색할 수 있도록 하고 있다.

이들 검색방법은 간단한 검색어뿐 아니라 Boolean검색과 같이 다양한 검색식을 제공하게 되어 사용자가 쉽게 원하는 정보를 찾을 수 있도록 도와주고 있다.

나. Full Text Search 의 특징

문서정보에 대해 검색을 하고자 한다면 기본적으로 문서의 유형, 저자, 제목 등의 찾고자 하는 문서에 대한 어느 정도의 정보를 사용자는 알고 있어야 한다. 또는 원하는 정보를 보유하고 있는 웹사이트를 찾고자 할 경우 웹사이트의 이름, URL 등을 알고 있어야 한다. Full text search에서는 찾고자 하는 정보의 핵심이 되는 주제어만 알고 있으면 검색이 충분하다. Full text search에서는 해당 정보에 대해 인덱싱을 하여 그 결과를 DB로 구축함으로써 사용자가 쉽고 빠르게 검색할 수 있게 된다.

다. Full Text Search 의 기술

(1) 저장

크게 모든 문서정보를 DB로 구축하는 경우와 문서에 대해 색인어를 DB로 구축하는 것으로 나눌 수 있다. 사용자가 원하는 문서에 대한 색인어를 이 색인어 DB를 통해 원문을 검색할 수 있게 한다.

(2) 자동 색인

자동색인은 사용자가 해당자료에 접근할 수 있도록 문서에서 자동으로 색인하여 항목을 생성하는 시스템이다. 이런 자동색을 위해 자연어처리, 형태소분석 및 인공지능적 기술이 사용되고 있다.

(3) 시소러스 **DB**

보다 효과적인 검색을 위하여 사용자가 표현하는 검색어와 문서에서 색인된 색인어의 의미상 언어통제를 해야 한다. 이때, 사용자의 검색어는 시소러스 용어사전을 참조하여 색인어DB를 탐색함으로써 검색의 정확성을 향상시키는 역할을 하게 된다.

(4) 검색

기본적으로 검색은 검색어를 통하여 이루어지며, 효율적인 검색기능을 제공하기 위해 Wild Card, Boolean 검색, 퍼지방식, 인공지능 방식 등을 지원하게 된다. Boolean방식이 가장 일반적으로 사용되고 있으며, 관련사전, 동의어 및 검색단어의 시소러스가 사용된다.

라. Full Text Search 의 적용분야

- 웹사이트 또는 웹사이트내의 HTML 문서정보 검색
- 기업내의 문서관리
- 전자도서관 서비스
- 연구소 및 각 공공기관의 기술문서 관리
- Intranet/Extranet
- 기타 방대한 양의 문서정보를 관리하는 곳 등등

기업내에서 문서정보를 공유하여 사용할 경우나 또는 특허청, 도서관 같은 많은 문서정보를 처리하는 곳에 실제로 구축되어 사용되고 있다. 또한, 엄청나게 증가하고 있는 웹사이트 및 웹상에서 정보들을 검색하기 위해 Alta

Vista와 같은 몇몇 검색엔진들은 Robot을 사용하여 Full text search를 제공하고 있다. 검색엔진의 경우는 사용자가 일일이 사이트를 다니면서 원하는 정보를 찾기란 거의 불가능하다고 할 수 있다. 이때, 검색 Robot이 주기적으로 인터넷상을 돌면서 색인어를 인덱싱을 하여 DB를 구축함으로써 사용자에게 편의를 제공하고 있다.

현재, Fulcrum, Verity, Basis+, BRS 등의 Full text search 시스템들이 소개되고 있으며, Fulcrum사의 경우 특허청, 도서관, 정부기관등의 실제로 적용되어 사용되고 있는 실정이다.

5. Parametric Search

가. Parametric Search 의 개요

Parametric search는 정보가 가지고 있는 속성에 따라 검색하는 것을 말한다. 즉, 같은 유형의 정보라 할지라도 그 성격에 따라 속성들이 달라지기 때문에 제한된 범위의 일정한 검색이 아닌 각 정보들만이 가지고 있는 특성에 따라 검색하는 것이다.

예를 들어, 인터넷 쇼핑몰에서 컴퓨터 주변기기를 구입하고자 할 때 스피커인 경우는 '제조회사, 판매가, 출력, 사용전압, 입력방식, 출력방식 등'을 고려해야 할것이며, 마우스인 경우는 '제조회사, 판매가, 방식(광, 볼) 등'을 고려하여야 할것이다. 즉, 같은 컴퓨터 주변기기에 대한 정확한 구매정보를 알기 위해서는 각 기기들이 가지게 되는 특성에 따라 검색해야 할 필요성이 요구된다.

위와 같이 서로 다른 속성들을 가진 정보들을 동일한 구조를 통해 검색하는 것보다는 각각의 특성에 대해 검색을 하는 것이 보다 효율적인 검색이 가능하게 된다. 또한, 이들 정보에 대해 일관되고 직관적인 분류체계를 부여하면 보다 쉽게 사용자들이 정확한 정보에 접근 할 수 있게 된다.

나. Parametric search 의 특징

Parametric search는 Object-oriented 기술을 이용하고 있다. 각 정보를 표현하기 위하여 Class와 Attribute를 사용하게 된다. 클래스는 정보를 나타내는 분류기준이며, 어트리뷰트는 각 클래스마다의 속성을 나타낸다. 정보를 검색하는 방법은 바로 이런 클래스와 어트리뷰트에 의해 이루어 진다. 즉, 찾고자 하는 정보에 대해 정보의 분류기준에 해당하는 클래스를 선택하고 그 클래스에 해당하는 상세한 parameter값을 입력하여 검색하게 되는 것이다.

클래스	어트리뷰트
상품정보	제조회사, 가격
의류	제조회사, 가격, 재질
남성의류	제조회사, 가격, 재질, ...
여성의류	제조회사, 가격, 재질, ...
유아의류	제조회사, 가격, 재질, ...
스포츠제품	제조회사, 가격
스포츠의류	제조회사, 가격, ...
신발	제조회사, 가격, 사이즈, 종류, 색상, ...
운동기구	제조회사, 가격, 종류, 색상, ...
....	...

예를들어 인터넷 쇼핑몰에서 신발에 대한 구매정보를 알아보려고 한다면 다음과 같다. 위의 표에서 처럼, 사용자는 신발을 찾기위해 먼저 스포츠용품을 선택하고 스포츠용품 중 신발류를 선택한다. 그 다음, 원하는 사이즈, 종류, 색상 등을 입력하여 검색을 하게 된다. 그러므로, 사용자는 찾고자하는 제품에 대해 정확한 지식 없이도 직관적으로 검색을 할 수 있게 된다.

특히, 클래스는 상속관계를 가지게 되므로 상위클래스는 일반적인 속성을 가지게 되고 하위의 클래스로 갈수록 보다 더 상세한 속성을 가지되는 것이다. 따라서, 사용자는 사람이 정보를 검색해 나가는 방식으로 검색이 가능하게 된다. 즉, 일반적인 정보에서부터 보다 상세한 정보를 찾아나가게 되는 것이다.

또한, 클래스를 살펴보면 같은 상품을 나타내고 있지만, 그 유형에 따라 속성값이 달라지게 된다. 그래서, 해당 정보만의 속성값에 의해 정보를 검색하게 되므로 보다 정확한 검색이 가능하게 된다.

다. Parametric Search 의 적용분야

- 웹사이트
- 문서관리
- 도면관리
- 전자도서관
- 부품정보관리, 공급자정보관리
- 인터넷 쇼핑몰
- Intranet/Extranet
- 기타 분류기준을 통해 검색이 가능한 정보 등등

분류기준을 통해 검색이 가능한 모든 정보에 대해 적용할 수 있다. 생산업체의 부품정보관리와 같이 적게는 수천건 많게는 수십만건에 해당하는 부품정보를 가지고 있으며 정확한 제품정보의 검색을 요구하는 곳에서 효율적으로 사용되고 있다. 다양하고 방대한 제품정보를 정확히 찾아내는 것은 제품코드 및 규격에 대한 정확한 지식 없이는 어렵다고 할 수 있다. 하지만, Parametric search를 통해 제품의 정확한 지식없이도 직관적으로 원하는 정보를 신속하게 찾아낼 수 있는 것이다. 이 외에도 문서, 도면, 전자도서관 등의 문헌, 도면 정보검색에 활용될 수 있으며, 특히, 인터넷 쇼핑몰에 대해 상품정보 검색에도 활용될 수 있다.

현재 CADIS사의 KrakatoA라는 Parametric search 시스템은 부품과 공급자에 대한 정보를 구축하여 효과적으로 사용되고 있으며, 전세계적으로 반도체에 대한 구매정보를 제공하고 있는 IHS와 같은 부품정보제공 서비스를 지원하고 있다.

6. 웹상에서의 효율적인 검색을 위한 방안

일반적으로 웹상에서의 검색은 Yahoo!, AltaVista 등과 같은 유명한 검색엔진이 사용되고 있으며, 가장 많이 활용되는 것이다. 이들 검색엔진들은 웹사이트내의 HTML 문서형태의 정보를 효율적으로 검색하고 있다. 하지만, 이제는 웹에서 HTML문서 뿐아니라 다양한 유형의 정보원들이 존재하고 있으며, 특히, 기업 내부의 DB로 구축된 정보들이 Web을 통하여 제공되고 있어, 데

이터의 유형 및 정보검색의 목적에 따라 검색이 이루어 져야 할것이다.

(1) 정보원의 유형별 정보검색

정보원의 유형에 따라 적절한 검색기법이 사용되어야 할 것이다. 내부에 구축되어 있는 DB를 검색하게 된다면 SQL을 통한 검색만으로 충분하지만, 이런 정보들이 웹상으로 서비스된다면 보다 적절한 다른 검색기법이 필요할 것이다.

현재 웹에서 일반적으로 사용되는 검색엔진들은 웹상에 존재하는 모든 정보를 검색할 수 있지만, 전문적인 DB를 통해 제공되는 정보의 검색에 대해서는 적당하지 않게 된다.

인트라넷을 중심으로 기업내의 문서관리가 이루어지고 문서정보의 공유를 위해 검색이 필요하다면 해당 문서가 존재하는 곳까지 검색엔진의 Keyword방식을 통해 찾아 갈수 있지만 실제 필요한 문서를 검색하기 위해서는 Full text search가 필요하게 된다. 또한, 웹상에서 부품/상품 또는 구매에 대한 정보가 제공된다면 사용자가 원하는 정보를 정확히 찾아내기 위해서는 정보의 다양한 속성에 의해 검색이 가능한 Parametric search를 통해 검색하는 것이 효과적이라고 할 수 있다.

이와 같이 다양한 유형의 정보에 대해 그 정보들이 가지고 성격에 따라 검색을 해야 할 것이다.

(2) 검색목적에 따른 정보검색

앞에서 기술했듯이 검색목적에 따라 정보검색이 몇가지로 분류된다.

전략검색과 같이 원하는 정보에 대해 포괄적인 정보를 검색하고자 한다면 검색어를 입력함으로써 정보를 찾아내는 Keyword 또는 Full text 검색이 유용할 것이다. 하지만, 이들의 검색결과는 사용자의 의도와 다르게 해당 검색어를 포함하는 수많은 결과를 가져오거나 또는 한건도 찾지 못하는 상황이 발생하게 된다. 이때, 해당 정보가 없는 것인지 아니면 검색식이 잘못된 것인지 불분명하게 된다.

만약 단순정보와 같이 정확한 정보검색을 요구할때는 정보가 가지고 있는 속성에 따라 검색을 하는 Parametric search가 유용하게 사용될 수 있다. 또한, 검색의 목적에 따라 혼합된 형태의 정보검색도 가능하다.

이와 같이 정보를 검색하는 목적에 따라 적절한 검색기법을 선택해야 할것이다.

(3) 정보 및 데이터들의 통합화

웹상에 존재하거나 웹과 연동이 되어 제공되는 정보에 대해 효과적이 검색기법을 통해 정보의 공유 및 활용하는 것도 중요하지만, 그 검색된 결과에 대해 타 정보시스템과의 통합화도 고려되어야 한다.

만약, 생산을 위해 전세계적으로 판매되는 반도체에 대한 구매정보를 검색하였다고 하면, 해당 반도체에 대해 제조회사, 모델명, 가격등에 대한 정보를 얻을 수 있다. 이때, 반도체에 대해 관리되고 있는 Databook이 제공된다거나, 제품의 사진 또는 환율이나 수입에 대한 정보를 연계 시킨다면 보다 효율적으로 검색된 정보를 활용할 수 있을 것이다.

즉, 방대한 양의 정보를 검색하여 원하는 결과를 얻을뿐 아니라, 검색결과와 관련된 정보 및 데이터에 대한 통합화를 제공한다면 정보검색의 효과를 극대화 시킬 수 있을 것이다.

7. 결론

정보화 시대를 맞이하여 새삼 정보의 중요성이 강조되고 있다. 이제는 데이터를 가공하여 유용한 정보를 생성하는 것도 중요하지만, 이런 정보의 폭발적인 증가에 의해 실제 사용자가 필요로 하는 정보를 활용할 수 있도록 하는 정보검색방법 또한 중요한 이슈로 대두되고 있다.

본 논문에서는 지금까지 소개되어 있는 몇가지 정보검색기법을 다음과 같이 고찰하여 보았다. RDB를 중심으로 기업 내부의 정보시스템에서 사용되는 SQL-Based 검색과 검색엔진 및 문서정보검색에 제공되는 Keyword, Full text 검색과 객체지향 개념을 도입하여 정보의 속성별로 검색을 하는 Parametric 검색기법에 대해 정의 및 특성들을 고찰하고, 이들 검색기법을 통해 웹상에서 보다 효율적인 검색을 위한 방안을 제시하여 보았다.

현재 인터넷이라는 거대한 네트웍 인프라를 기반으로 웹을 통해 엄청난 양의 정보들이 제공되고 있다. 이제는 기존의 데이터에서 정보를 가공하는 수준에서 벗어나 실제로 사용자가 원하고 필요한 정보를 적절히 활용할 수 있도록 정보를 'Knowledg'의 수준까지 제공해야 하는 필요성을 가지게 되었다. 그러기 위해 웹상에서 적절한 정보검색기법의 활용이 요구되어야 할 것이며, 보다 효과적인 정보검색기법을 고려하여야 할 것이다.

[참고문헌]

인터넷 정보검색의 정의와 분류법, www.searchplus.com/bution/contri.html
정보검색방법론, www.withnet.co.kr/~major/search/pre.html
김수천외 1명, 정보검색의 이론과 실제, 바다출판사, 1997.
김화수, 인터넷 정보검색의 마지막 노하우, 바다출판사, 1997.
김휘출, 정보검색과 인터넷, 홍익미디어, 1997.
정보검색이란, http://www.three.co.kr/menu_retrieval.html