



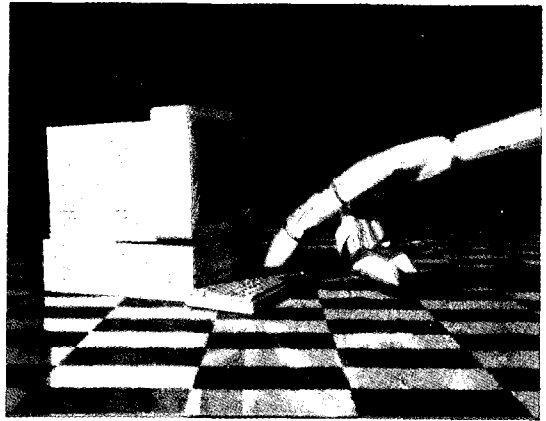
데이터베이스 정보 검색 기법

이용호 / 데이콤
부가통신사업본부 IP 영업부 과장

I. 서언

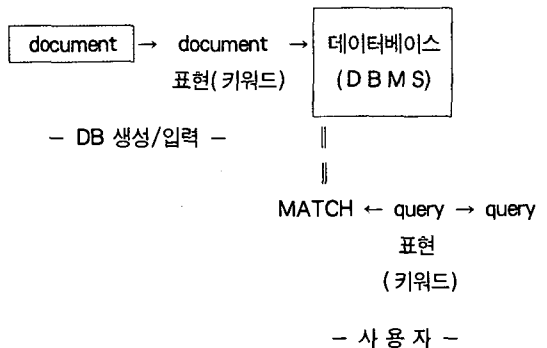
우 리나라 데이터베이스 산업의 취약을 만회하기 위한 民·官의 노력의 일환으로 국가 공공 데이터베이스를 개발이 구체화되고 있는 가운데, 대형 DB등장으로 예상되는 정보검색 이용 방법의 다양화/다변화로 인한 이용자 혼돈이 예상되고 있다.

해외 DB인 DIALOG, LEXIS/NEXIS에서 보는 바와 같이 외국의 경우도 상당한 기간이 경과한 후 검색 시스템이 체계화되어 지는 것이다. 검색 시스템 체계의 정례화는 곧 데이터베이스의 유통 및 이용 확산에 필수적인 필요 조건이며 정보를 쉽고 경제적으로 찾을 수 있는 첩경이다. 따라서 본고에서는 일반적 DB 검색이론과 현재 유통되고 있는 키워드(KEYWORD) 검색 기법을 소개하고자 한다.



인(INDEX)형태로 표시하고 사용자의 요구에 맞는 절차가 수행되는 것이다. 한 예로 책을제목이나 저자명으로 표시해둔 색인함에서 사용자가 책 이름이나 저자명을 완전히 알지못하더라도 도서관에서 서지목록으로 冊을 찾아볼 수 있는 것과

[그림 A: 사용자의 질의에 의한 "Best matching"의 개념]



II. 검색 시스템의 이론 (KEYWORD 시스템)

1. 키워드 검색 시스템의 개념

키워드 검색이란 사용자 질의어에 의한 표현을 입력하여 컴퓨터에 저장된 데이터베이스 자료와의 추출처리가 일어나 추출(Match)된 자료 결과를 통하여 가장필요 적절한 정보를 질의 할 수 있는 기능을 말한다. 이때 대상들은 키워드나 색

같은 개념인 것이다. <그림 A>는 사용자의 질의에 의한 최적 검색(Best Match)의 개념을 표현한 것이다.

키워드 검색 방법이 보편화된 근래와서 널리 사용되나 50% 정도의 적중률 보이고 있어 완벽하지는 않다는 단점이 있다. 이는 한 단어가 여러 가지의 뜻을 가질 수 있고 같은 의미로 여러 가지의 말이 사용가능하기 때문에 생기는 어찌면 당연 문제일 것이다.

2 키워드 검색의 필요성

키워드검색 대상들이 다양해 지면서 기존의 메뉴 구조의 순차적 조회 방법들이 너무 비효율적인 관계로 키워드에 근거한 정보의 검색 시스템이 점차 많아지고 있는 실정이다.

그리하여 정보처리 분야(information science)나 문헌정보 검색분야에서는 검색 효율을 높이기 위한 갖가지 노력을 경주하고 있다. 특히 관계형 데이터베이스가 강세를 보임으로써 데이터관계들의 구조가 다양해지고는 있지만 아직 처리가 매끄럽지 못한 불변성(inflexible)과 구조적 제한성(limited)때문에 키워드 방식의 검색은 꼭 필요하게 되는 것이다. 근래와서 정보통신 기술의 발달로 인한 지협적 DB 사용보다 광대역 DB사용이 필수적인 요소로 등장하게 된 것 또한 키워드 검색이 필요한 원인이 되기도 한다.

3 키워드 검색 시스템의 효율

검색 시스템 효율분석이나 성능 평가(measure performance)를 하기는 상당히 어려운 문제이긴 하나 시스템이 처리결과로 돌려주는 자료(문서)들의 관련 비율 즉, 특정한 질의에 대해서 관련된 모든 자료의 검색(all and only) 여부에 달려 있다고 볼수있다. 이상적인 검색시스템이란 질의

내용의 단순화에도 불구하고 관련된 자료 추출이 정확하고 빠르게 검색될 수 있는 알고리즘이 있어야 하는 것이다.

검색 결과에 대한 평가 기준은 검색 시스템이 논리적 절차에 따라 검색되는 1차산출물(Thoughtput)의 비율을 효율(Recall)이라고 하며, 1차 검색된 자료중에서 찾고자하는 항목과 관련된 자료(Record건수)가 실제로 존재하는 비율을 정밀도(Precision)라고 한다.

이상적인 검색시스템에서는 높은 효율, 높은 정밀도(Precision)를 유지하는 것이 목표이나 실제의 경우 많은 데이터베이스에서는 효율(Recall)과 정밀도(Precision)는 교환관계로 효율(Recall)이 높으면 정밀도(Precision)가 작아지고, 정밀도(Precision)가 크면 효율(Recall)이 작아지는 현상을 보이는 것이다.

효율(Recall)과 정밀도(Precision)을 증가시키는 방법으로는 다음과 같은 방법이 있을 수 있다.

가. 효율(Recall)을 증가시키는 기법

- 접두/접미를 이용한 어구생략(Term truncation)을 이용
- 동의어(synonym)의 첨가
- 관계된 품의 사용
- 확장된 품m의 사용

나. 정밀도(Precision)을 증가시키는 방법

- Term weighting
- 구(Phrase)나 중분(Compound)의 추가
- 협의의 단어로 치환/대치

키워드 검색의 문제에 대한 심리학적인 분석에 의하면 하나의 검색대상을 접근하는데 유사의 명칭(Unlimited alias)을 사용할 수 있어야 효율(Recall)의 비율이 증가하는 것으로 나타났다. 그러나, 여러개의 명칭에 혼용해서 사용함으로써 검색된 자료의 모호성이 발생할 확율이 높을 수도 있다.

4. 주요한 키워드 검색 용어

○ 잠재적 語義 색인구조(Latent Semantic indexing)

사람이 사용하는 말은 동의어(Synonym)와 多義性(Polysemy)이 있어서 직접적인 언어합성(Term-matching)은 상당한 비효율성을 야기할 수 있다.

정보를 찾을때 어의론적 의미에 기반을 두고 색인구조를 갖는 것을 잠재적어의 색인구조(Latent Semantic indexing)이라 한다.

○ 연관형 색인구조 (Related indexing ideas)

언어의 다양성을 극복하기 위한 노력으로 사용 가능한 어휘의 제한하면 좋은 성과를 얻지는 못함으로 동의어 사전들의 사용하여 관련된 언어(term)들을 클래스(Class)로 묶어 사용함으로써 효율(recall)을 증가시킬 수 있도록 만든 구조를 말한다. 이렇게 형성된 색인구조는 검색 효율의 증가 시킬 수가 있다.

○ 사용자요구 충족 검색 방법 (Relevance Feedback)

초기의 검색방법은 보통 배치(BATCH) 모드에서 수행하여 한번 수행된 결과에 만족할수밖에 없었으나 최근 대부분의 데이터베이스 검색시스템에서는 대화형(interactive) 시스템 환경으로 변화되었기 때문에 사용자로부터 즉각적인 충족이 필요하다는 뜻, 즉 요구충족 검색(Relevance Feedback)의 개념은 어떤 아이템이 사용자의 관심이나 재조건 대상이 되는가에 따라 질의의 표현을 바꾸어 재검색할 수 있는 기능을 말한다.

다음은 일반적인 요구충족 검색(relevance feedback)의 수행절차이다.

● 수행 순서

1. 사용자가 질의를 줌
2. 시스템이 후보레코드(candidate들)을 돌려줌
3. 사용자가 어떤 아이템이 가장 관련되는가를 판단
4. 이를 사용해서 초기의 질의를 수정/보완 추가조건입력
5. 최적 관련 항목도출 및 조회(검색효율 상승)

5. 검색 데이터 형식과 차이점

검색할 데이터 형식의 종류와 특징은 <표 1>에서 보는 바와같이 크게 정형 데이터와 비정형 데이터로 구분되어 진다.

<표 1>

데이터형식	특 징
정형 데이터 (Formatted data)	<ul style="list-style-type: none"> - 잘 정의된 형식(well-defined) - 논리적인 구조(logical organization) - 사용자가 형식을 알고 있음 - 유일한 의미(unique meaning) - 여러 사용자가 같은방법으로 접근 가능 - 효율적인 정보검색 가능 제공
비정형 데이터 (Unformatted data)	<ul style="list-style-type: none"> - 일정한 구조가 없음 - 대부분의 새로운 응용에서 사용 - 복잡한 정보검색 기법 필요 - 기존 시스템의 기능 확장 필요

데이터 구조(Data Structure) 측면에서 볼때 정형 데이터인 경우에는 관계형 데이터베이스의 경우 일정한 수의 애트리뷰트(ATTRIBUTE)의 값은 하나의 애트리뷰트에 대응하여 검색 기법이 많고 비정형 데이터의 경우 문서형(document)인 경우 임의의 수의 단어로 구성되어 있어 대응되는 애트리뷰트가 거의 존재하지 않는 것이 특징이다.

따라서 완전한 검색을 위해서는 전체 자료 검

색을 수반할 수 밖에 없다. 각 문서의 효율적인 검색을 위해서 키워드를 추출하는 인덱싱(indexing)에 있어서는 각 키워드를 애트리뷰트로 여길 수 있으므로 정형 데이터 경우와 같이 유사성이 증가하게 할 수 있다. 그러나, 키워드의 양은 가변이라는 차이점이 있다.

6. 질의의 타입(Types of Query) 과 특징(Feature)

질의의 형태	특 징
● 부울린(boolean) 대수에의한 질의	- 사용자가 확신을 가지고 질의하는 경우에는 유용 - 실제 검색상황과는 배치성 높음 - 관련된 항목을 찾는 것과 같은
● "fuzzy" 이 론 형 질의	- 확실하지 않은 질의 - 정보검색 기법의 연구에 있어서 중요한 이슈
● 가중 부울린 연산자 (weighted Boolean operator)	- 인간의 사고/판단에 근접 - 부울형과 퍼지형을 혼합하여 질의하는 방법 - 위의 두 방법의 개선 - 실험적 효율은 높은 편임

III. 검색기법의 종류와 특징

1. 다중 키 검색 기법(MULTI-KEY Method)

하나의 기본 키 대신에 여러 개의 보조키(Secondary Key)를 사용해서 레코드를 검색하는 방법으로 가장 전통적인 검색방법으로 가장 많이 사용하고 있다. 논리곱(AND)이나 논리합(OR)의 표현이 가장 일반적인 형태이나 텍스트의 검색에는 비교적 부적합한 검색방법

2 역 화일 방법(Inversion Method)

최근 자료는 파일의 끝 부분에 있다는 데 착안하여 기존 다중 키 검색방법의 단점인 탐색 공간

을 많이 줄임 다중키의 경우 많은 양의 저장 공간과 검색에 필요한 버퍼(BUFFER)가 필요: 원래 화일의 50~300%를 사용

3. 다중 속성 해싱 기법(Multi-Attribute Hashing Method)

각각의 애트리뷰트(Attributes)를 해쉬(hash)해서 일정한 수의 비트가 세트된 패턴으로 바꿔 이들에 대해 슈퍼임포우징(superimposing)이나 칸케터넷팅(concatenating)기법을 사용해서 레코드 시그니춰 생성하는 기법

※ 슈퍼임포우징(Superimposing) : 애트리뷰트의 구별이 없는 간접키 검색 방법의 일종 텍스트의 검색에 가장 적합

4. 시그니춰(Signature) 파일 방법

단순한 화일 구조를 갖고 있는 경우 적은 저장 공간의 사용하여 레코드의 검색을 하며 사용하는 인덱스 언어(term)들의 추상화도 가능하다.

각 아이템의 존재 여부에 대한 지표가 판단의 기준이 되며 각 레코드에 대해서 대응시킨 결과를 검색하는 방법 원래 화일의 10~20%만을 사용하므로 대용량의 데이터베이스에 적용하는 것이 바람직함. 적용분야로는 텍스트 검색(text retrieval), 사무 시스템(office systems), 프로로그 데이터베이스(prolog database) 등이 있다.

다음은 시그니춰(SIGNATURE) 질의의 처리하는 절차를 나타낸 것이다.

1. 데이터 화일 대신에 먼저 시그니춰 화일을 볼(시그니춰 화일이 관련되지 않는 부분을 필터링(filtering))
2. 찾고자 하는 아이템이 있는가 결정
3. 찾고자 하는 아이템이 있는 일부분만 검색

5. 전체 텍스트(FULL TEXT) 검색

문서를 순차적으로 읽어나가면서 주어진 문자열을 포함하고 있는가를 알아내는 가장 직접적인 방법으로 전체 텍스트 검색이 있다. 전체 텍스트 검색방법으로 일반적인 정규 수식(Regular expre-



ssion)의 사용하는 방법과 주어진 식이 맞는가를 알아내기 위한 유한 문자구성(Finite automata)하는 방법이 있다.

문자열을 탐색할때는 불필요한 문자열(don't care character)을 포함하지 않는 문자의 순차적 검색으로 이루어진다. 전체 텍스트 검색에 있어서 질의 결정(query resolution) 단계의 필요한데

찾고자 하는 질의의 사항이 단순하지 않은 경우에 알맞는 항목을 찾았는가의 여부를 결정하기 위해서는 부울린 수식의 논리적 검사가 필요하다. 탐색 시간 문서의 크기에 비례한다. 현재 서비스되고 있는 대표적인 전체 텍스트(Full Text) 검색으로 뉴스 크리핑 서비스가 있다.

6. 逆順 색인 기법(Inverted Index Method)

레코드(또는 텍스트)의 단어를 모두 인덱스로 만드는 경우를 가정하여 각 레코드에 대해 색인 작업을 하여 전체화일을 레코드나 텍스트의 배열로 표현하는 방법 배열의 행은 한 레코드에 포함된 키워드의 집합이며 레코드-키워드 배열을 뒤집어 즉, 배열의 행이 한 키워드를 포함하는 레코드들이 되도록 구성하는 방식으로 입력된 질의에 의해 검색을 수행하다.

○ 역 색인 검색 과정의 장단점

역 인덱스 검색 과정의 장점은 리스트 머징(List Merging) 연산으로 쉽게 구현한다는 점이고 실제 레코드는 최종적으로 검색함으로 속도가 비교적 빠르다는 점이다. 또한 동의어(Synonyms)를 쉽게 지원할 수 있는 장점이 있다. 단점으로는 탐색 공간의 오버헤드가 많고 인덱스를 갱신하고 재구성하는데 드는 비용이 비교적 많이 든다는 점이다(같은 단어의 나열을 합하는데 드는 비용)

▶ FULL TEXT 검색의 장단점

장 점	단 점
<ul style="list-style-type: none"> ● 여분의 공간이 불필요 ● 삽입(INSERT)과 삭제(DELETE)에 있어서 노력이 불필요 	<ul style="list-style-type: none"> ● 응답시간이 늦음 ● 대용량의 데이터베이스의 경우에 심각한 문제 <ul style="list-style-type: none"> - 개선된 알고리즘의 사용 - 특정 목적의 H/W를 사용해 개선

연산자	검 색 방 법	표 현 형 식
OR	두 키워드 중 하나 이상이 포함된 레코드나 텍스트를 검색	키워드 1 OR 키워드 2
AND	두 키워드가 모두 포함된 레코드나 텍스트를 검색	키워드 1 AND 키워드 2
NOT	키워드가 포함되지 않은 레코드나 텍스트를 검색 보통은 AND 와 함께 사용하는 것이 보통	NOT 키워드

IV. 검색 효율(RECALL)을 높이기 위한 검색기법 소개

1. 부울린(BOOLEAN) 질의의 기본 요소

찾으려는 키워드들에 대해 부울린 연산자 AND, OR, NOT을 사용하여 부울린 수식으로 표현되는 연산자로 아래 같은 종류가 있다.

각 연산자에 대해 단계 별로 리스트 머지(list merging)과정을 거쳐 수행된다.

각 리스트는 레코드 순서대로 정렬되어 있다고 가정하면 부울린(BOOLEAN) 연산자가 없는 벡터 질의의 경우는 각 키워드에 해당하는 레코드 리스트를 모두 합한 것(Si, Sj, Sk)으로 표시한다. 부울린 연산의 경우별 처리 형태는

- 부울린 AND 질의의 경우 : (Si and Sj)
 - 두 언어의 리스트를 더해 중복된 레코드들만 추출(교집합)
- 부울린 OR 질의의 경우 : (Si or Sj)

- 두 언어의 리스트를 더해 모든 레코드를 중복되지 않도록 취함 (합집합)

- 부울린 NOT 질의 (AND 연산자와 함께 사용되는 경우) : (Si and not Sj)
 - (Si and not Sj)의 조건을 충족하는 리스트를 구하고 Si의 리스트와 합해 중복되지 않는 레코드 선택

적용예) 질의(Query) : ((Si OR S2) AND NOT S3) 이라면 (INDEX 리스트 참조)

2 검색질의어의 조건 확장 (동의어 지정/추가)

정의한 키워드에 대해 동의어를 추가로 보완 유지하여 질의시 동의어를 자동적으로 첨가시켜 검색 범위를 넓히는 방법(효율 증가가 목적)

- 부울린 질의의 경우 OR 연산자로 동의어를 포함시킴

<INDEX 리스트>

INDEX 리스트	S1 : {R1, R3}
	S2 : {R1, R2}
	S3 : {R2, R3, R4}

처리과정	(S1, S2) 더함 : {R1, R1, R2, R3}
	(S1 or S2) 출력 : {R1, R2, R3}
	((S1 or S2), S3) 더함 : {R1, R2, R2, R3, R3, R4}
	((S1 or S2) and S3) 출력 : {R2, R3}
	((S1 or S2), ((S1 or S2) and S3)) 더함 : {R1, R2, R2, R3, R4}
결 과	((((S1 or S2) and not S3) 출력 : {R1, R2, R2, R3, R3, R4}

- 원래 질의 : ((T1 and T2) or T3)
- 동의어 : T1 - S1, T3 - S3
- 변환된 질의 : ((T1 or S1) and T2) or (T3 or S3)

3. 語句 생략 (Term Truncation)

질의어의 키워드 일부만을 지정하여 검색 범위를 넓히기 위하여 검색조건을 생략하는 방법 (호을 증가)

- psych? : psychiatrist, psychiatry, psychiatric, psychology, ... 등(?는 가변 길이의 대치 문자열을 나타냄)

그러나 단어, 어절의 일부는 검색에 중요하지 않을 수도 있지만 우리말의 조사 어미등을 무조건 생략하는 것은 엉뚱한 결과를 초래할 가능성이 있다.

[語句 생략(Term Truncation) 구현의 표현의 예]

- 접미어 생략(suffix truncation) : 예) psych?
- 접두어 생략(prefix truncation) : 예) ?symmetry
- 접두어와 접미어 동시 생략 : 예) ?symm?
- 내부어구 생략(infix truncation) : 예) wom?n

V. 천리안을 통한 실제 검색의 사례 (인물정보 중심)

인물정보(PEOPLE) [정보검색]	천리안 제공
1. 인물정보 검색	[정보안내] @ 11. 정보 안내 @ 12. 정보 검색방법 및 출력 예 @ 13. 키워드 검색 안내
이 정보 이용료는 분당 NNN원 입니다 (@표시 정보는 분당 15원)	
초기메뉴 (T) 상위메뉴 (M) 직접이동 (GO) 도움말(H) 접속종료 (X, BYE) 기타 (Z) 선택> 1	

인물정보(PEOPLE)	중앙일보사(02-7515-451)
찾고자하는 자료와 관련된 단어나 검색식을 입력하십시오. 검색항목번호를 이용하여 검색하고자 할 경우 항목번호=검색어 형태로 입력하십시오.	
항목명	번호 사 용 예
· 성명, 기관 및 직위, 별칭, 주소, 학력, 경력, 저서, 논문, 취미, 특기 등	1 (1=)인천 * 교수 * 경역학?
· 성명	2 2= 홍길?
· 기관 및 직위	3 3=데이콤 * 3=전무?
· 출신학교	4 4=고려대?
· 본적/출생지	5 5=충남
예) 경기도 출신의 국회의원을 찾고자 한다면 검색식 ==> 경기? * 국회의원	
도움말 (H) 상위메뉴(M) 관련단어보기(E 검색어) 검색항목열람(D) 검색식출력(EX) 검색식 ==> 1=인천 * 교수 * 경영학?	
..... 지금 검색 중입니다.	
검색을 중단하실려면 Ctrl-C를 누르시오.	

지금까지 검색한 결과는 아래와 같습니다!!
 번호 건수 검색식
 결과1 15 1=인천 * 교수 * 경영학?

아래 메뉴를 선택하십시오!!!

1. 검색된 결과의 제목 및 내용을 보고자 할 경우 (L 결과번호)
2. 최종검색결과와 AND 연산을 하고자 할 경우
3. 최종검색결과와 OR 연산을 하고자 할 경우
4. 특정 결과번호부터 검색을 다시 시작하고자 할 경우 (C 결과번호)
5. 검색을 처음부터 다시 시작하고자 할 경우 (S 또는 Q)
6. 명령어 방식으로 검색으로 계속하고자 하는 경우

검색(번호) 도움말(H) 직접이동(GO) 초기메뉴(T) 상위메뉴(M) 기타(Z) 종료(X)
 선택 > 1. ※ (검색된 결과 리스트를 화면 출력 시키는 명령어)

인물정보(PEOPLE) 현재시각 : 94/04/21 18:11:53 자료량 : 15건
 검색식 : : 1=인천 * 교수 * 경영학?

번호	성명	기관	직위	생년월일
1	김기명(金基明)	인하대	교수	1943. 09. 24
2	신용휘(辛容輝)	인하대	교수	1942. 09. 14
3	신현주(申鉉柱)	세종대	교수	1939. 02. 05
4	전용수(全鎔秀)	인하대	교수	1943. 11. 02
5	정수일(鄭秀一)	인하대	교수	1943. 09. 18
6	노순규(魯淳圭)	한국기업경영연구원	원장	1955. 02. 05
7	이성주(李成柱)	인하대	교수	1944. 11. 20
8	이경환(李慶煥)	인하대	교수	1949. 02. 03
9	정재훈(鄭在勳)	인하대	교수	1951. 11. 01
10	천명섭(千命燮)	인하대	교수	1944. 08. 21
11	박기찬(朴基贊)	인하대	교수	1955. 07. 03
12	안용근(安勇根)	인하대	교수	1935. 05. 07
13	배무언(裴武彦)	인하대	교수	1949. 07. 17
14	이영찬(李泳燦)	인천대	교수	1937. 05. 10

검색(번호) 도움말(H) 직접이동(GO) 초기메뉴(T) 상위메뉴(M) 기타(Z) 종료(X) 목록
 열림(L) 연속출력(NR) 목록순서변경(SORT) FAX출력(FAX)
 선택 > 1 ※ (1의 세부 RECORD 항목이 DISPLAY 된다. 이하 생략)...

I. 결 언

지금까지 일반적인 검색시스템의 개념과 기법 그리고 천리안의 검색시스템을 통한 키워드 검색 실험을 살펴보았다. 공중 VAN 사업자의 증가로 정보 저장 형태도 정형 데이터에서 비정형 데이터로 바뀌어지고 있으며 이용형태 또한 키워드 방식에서 자동 색인어 방식이나 자연어 처리 방식으로 바뀌고 있는 추세이다. 새로운 응용분야

이기는 하나 최근 멀티미디어 등 여러 다중매체의 혼합에 의한 새로운 형태의 데이터베이스의 등장으로 검색방법도 기존 검색이론과 다른 신기술을 도입한 방법으로서의 연구가 지속적으로 이루어져야 할 것으로 본다. 데이터베이스 검색 분야에서 새로운 기법의 정보검색 기능과 새로운 DBMS (DataBase Management System)이 서서히 출현하고 있는 것이다.