WWW 에서 데이터베이스와 검색엔진의 연동을 통한 SGML 검색시스템의 구현

Implementation of SGML Retrieval System through Interoperability with Database and Search Engine based on WWW

김낙현, 정수용, 노명호
Nak-Hyun Kim, Su-Young Chung, Myeong-Ho Roh
Korea Digital Line
Seoul, Korea

Abstract

The advent of the Internet and the enormous increase in volume of electronically stored information (SGML, Image, Sound, etc.) has led to substantial work on IR(Information Retrieval). To service on the WWW, construction and retrieval technology of SGML, which is the fundamental standard data format for CALS/EC, is needed specially.

Due to such a change, it becomes essential to change the existing paradigm of conventional information retrieval systems and to adopt new Internet service system with search engine, SGML browser and advanced Internet technology on WWW. KIPRIS(Korea Industrial Property Rights Information Service), which is the specialized and integrated Internet service systems in the field of industrial property rights information service, is trying to be a guide for our country to establish its technological competitiveness with providing the online service of high quality.

The objective of the paper identifies features and technologies of KIPRIS IR(Information Retrieval) system based on WWW as follows. First, it describes the development background and process of KIPRIS. Second, it presents a fundamental technology that consists of IR(Information Retrieval) concept, BRS(Bibliographical Retrieval System) search engine, SGML implementation technologies and the Internet/WWW technologies. Third, it provides information about system configuration, architecture, and the features and characteristics of KIPRIS. Finally, the implemented KIPRIS system is introduced.

WWW에서 데이터베이스와 검색엔진의 연동을 통한 SGML검색시스템의 구현

Implementation of SGML Retrieval System through Interoperability with Database and Search Engine based on WWW

1999.7.13 'CALS/EC KOREA 국제학술대회
㈜ 한국디지탈라인 [www.kdline.co.kr]
김낙현, 정수용, 노명호

kdl

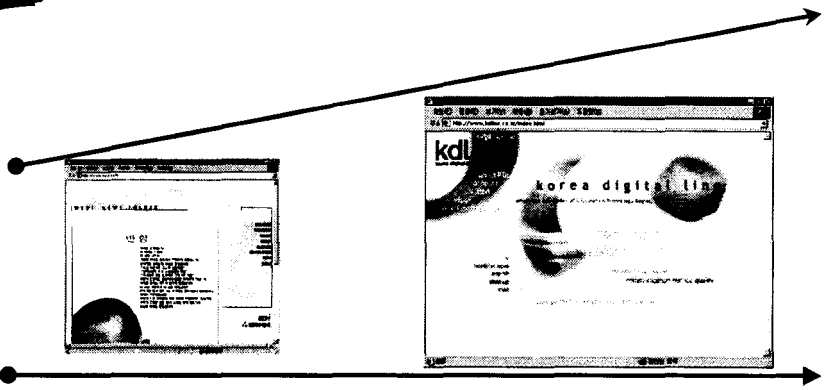
Content

- System
  - System Overview
  - System Objective
  - System Function
  - System Architecture
  - System Configuration
- SGML
  - SGML on the WWW
  - SGML on the Search Engine
  - SGML Construction and Retrieval
- Interoperability with WWW/Database/Search Engine

The advent of the Internet and the enormous increase in volume of electronically stored information (SGML, Image, Sound, etc.) has led to substantial work on IR(Information Retrieval). To service on the WWW, construction and retrieval technology of SGML, which is the fundamental standard data format for CALS/EC, is needed specially.

kdl

Korea Digital Line



1994.2 WebInternational

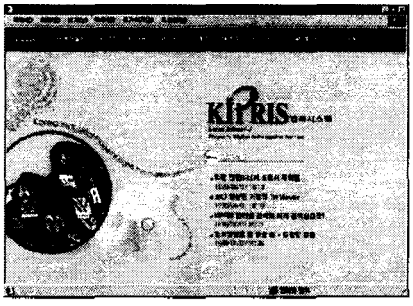
(Internet, Intranet, Extranet)

1998.9 Korea Digital Line

(Web based SI Service, Intranet Groupware, CALS/EC/EDI, SGML/XML/IETM/RM)

kdl 주)한국디지털라인

System Overview



- 특허기술정보 인터넷 서비스 시스템
  - 1997 정보화 지원사업(한국전산원)
  - 1997.12.1 - 1998.8.31
  - www.kipris.or.kr
  - WWW의 저변확대에 따른 인터넷 검색서비스 구축
  - 행정전산화 7개년 계획에 따른 SGML포맷의 데이터 제공
  - 대용량 데이터베이스를 위한 검색시스템 도입

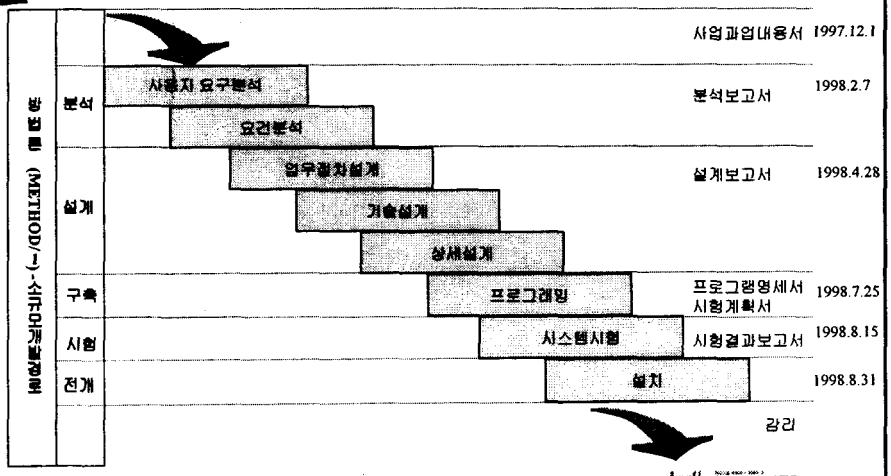
kdl 주)한국디지털라인

System Objective

- WWW을 이용한 **특허기술정보의 신속한 제공**
  - **효율적인 인터넷 검색 기법과 첨단 인터넷 기술 도입**
    - ▼ BRS 검색엔진도입(가중치 부여, 근접/인접 검색, 복합명사 처리 등)
    - ▼ 검색결과를 E-Mail 로 제공 (Mailing List), Plug-Ins 기술응용
  - **국제 표준 문서인 SGML 지원**
    - ▼ 신규 SGML DB구축 및 검색시스템 개발
  - **KIPRIS 서비스 시스템(C/S)의 보완**
    - ▼ 현 KIPRIC 특허정보 시스템의 서비스 환경 개선

kdl 한양국디지털라인

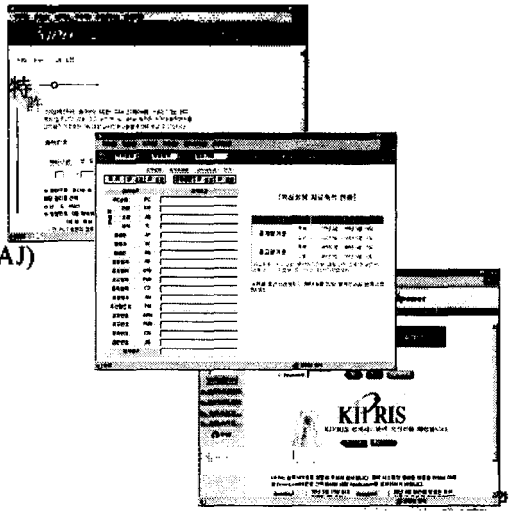
Project Schedule



kdl 한양국디지털라인

System Function

- 검색시스템
  - 특허실용
  - 의장
  - 상표
  - 심판
  - 해외(IFD, FPDB, PAJ)
  - 통계
  - 민원안내
- 홈페이지



System Characteristics

- IR (Information Retrieval)
  - Search Engine
  - Large Data Retrieval
  - Multimedia Data Retrieval
- Interoperability with
  - Web Technology
  - SGML/XML Construction and Retrieval
  - Search Engine and Database
  - Knowledge for Patent Information
  - Viewer(SGML, Tif, Statistics)

DBMS vs. IRS

- formatted data
- exact matching
- deterministic
- table
- small to large

- unformatted
- fuzzy
- probabilistic
- document
- small to large
  - generally IR deals with very large data

kdl

Patent Information

구 분	내 용	자료건수
특허/실용	서지사항, 중간처리정보, 대표도면, 초록(청구범위) 등록정보, 심판정보, SGML	1,597,671
의장	서지사항, 중간처리정보, 등록정보, 심판정보, 의장공보	193,808
상표	서지사항, 상표구성, 등록정보, 심판정보, 상표공보	1,512,461
심판	서지사항, 심판결문 정보	102,611
해외	특허영세서, 서지사항, 영문초록, 대표도면, SGML	13,895,895
	98년 8월	17,302,446

Text: 100GB
Image: 1.2 TB
27,000,000 건
Search Speed is the most CSF

kdl

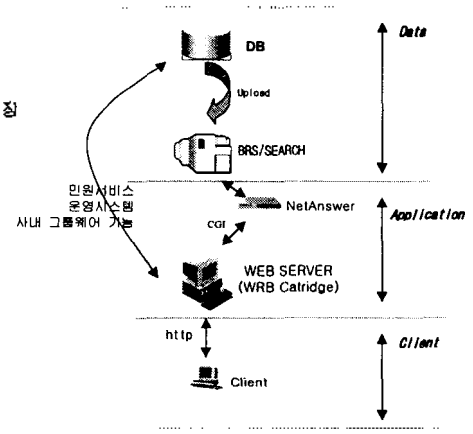
System Explanation

Web Server	Oracle Web Server 3.0 (HP 9000 D310)
Firewall Server	CheckPoint FireWall-1 (HP 9000 D210)
Search Engine	BRS/NetAnswer
Database	Oracle 8.04 (SGML, TEXT, IMAGE)
	HP-UX, HP9000 11대 with FileNet, EMC
	Interoperability with WWW and DB: PRO*C, PL/SQL

kdl 주)한국디지털라인

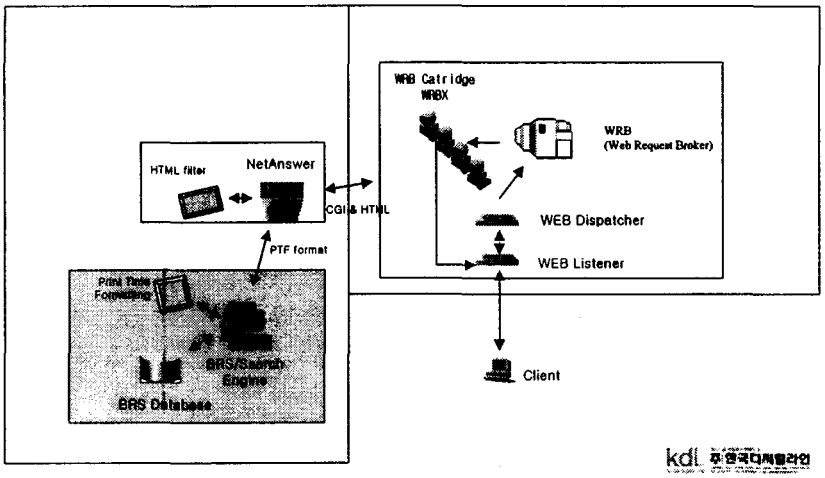
System Architecture(1)

- 3-Tier Architecture
  - DB Server
    - DB는 본래기능인 Master Data 저장
  - BRS 검색엔진
    - 검색부분 전담
  - NetAnswer
    - BRS와 Web Server의 상호연동
  - Web Application Server
    - 속도향상을 위한 방안
    - WRB Architecture로 구현
  - Client
    - 쉬운 사용자 인터페이스



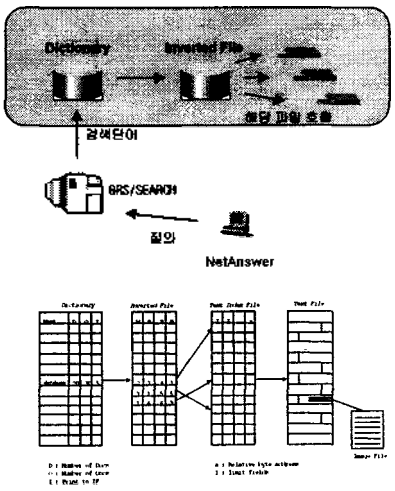
kdl 주)한국디지털라인

System Architecture(2)

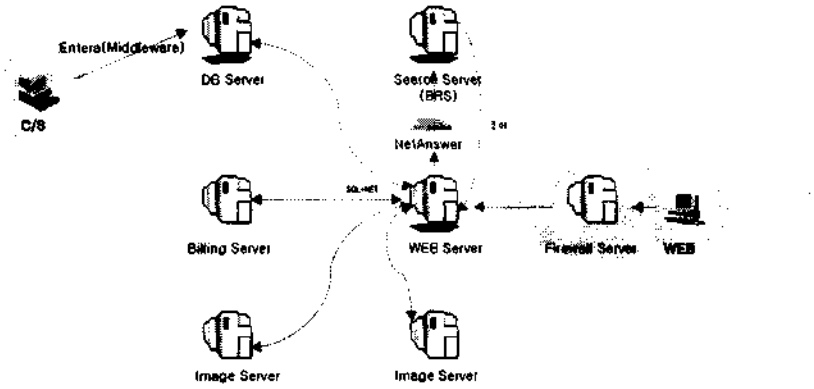


BRS/SEARCH

- BRS의 동작원리
  - BRS는 한글/영문 형태소를 통해 본문 중에서 색인어로 가치가 있는 단어만을 추출하여 Dictionary를 구성
  - 각 단어의 위치정보를 Inverted File에 저장
  - Inverted File에서 위치정보를 취하여 본문의 Index File을 통해 본문의 내용을 출력



System Configuration



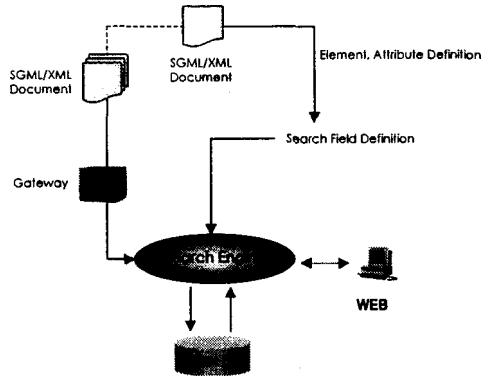
kdl

SGML on the WWW

- 필요성
  - 전자출원 및 특허행정업무의 전산화가 SGML 포맷으로 개발 중
  - SGML 자료를 인터넷을 통해 서비스
- 제공방안
  - SGML Plug-In
  - SGML Converter to XML
  - SGML Image Extractor
  - SGML Help Application

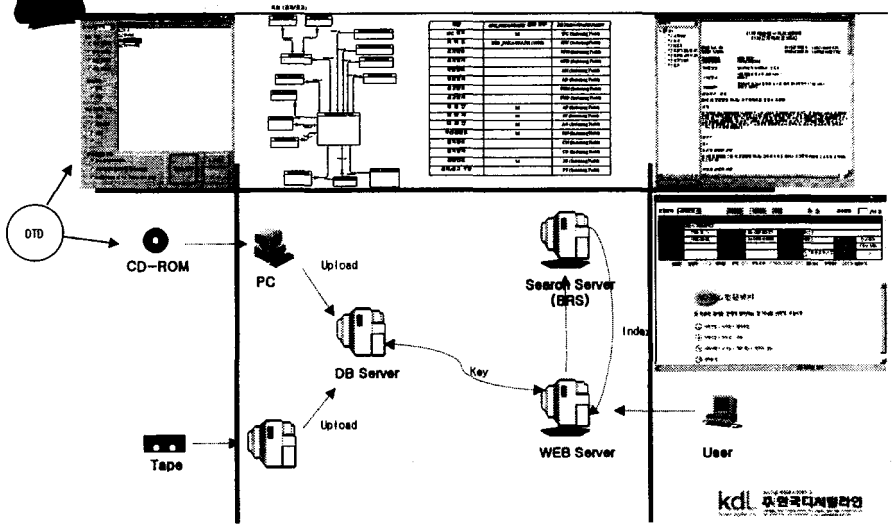
kdl

SGML on Search Engine



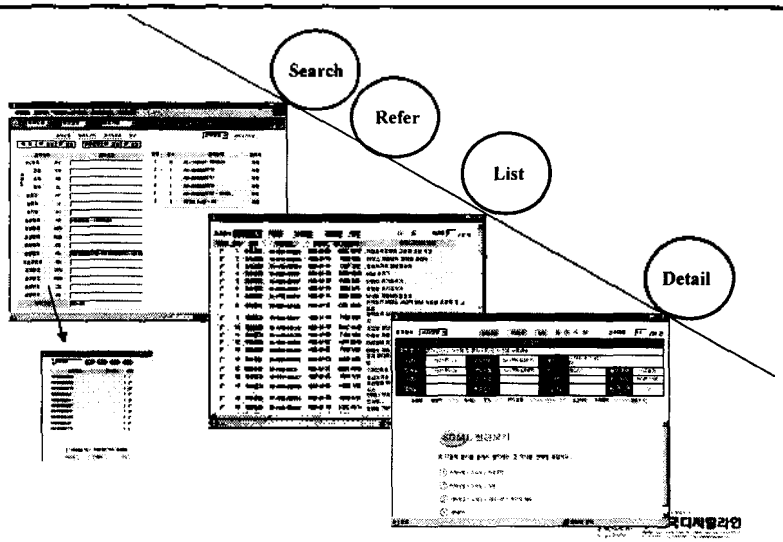
kdl

SGML Construction and Retrieval



kdl

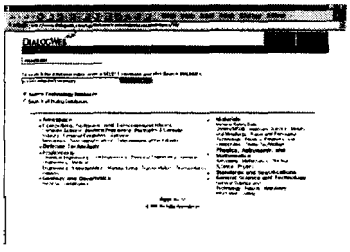
Search Example(SGML)



Comparison with DIALOG

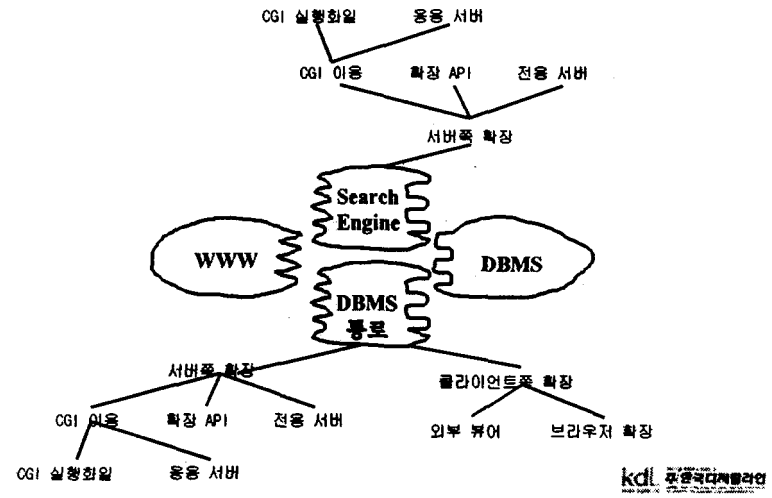
		질의검색	시간	기간검색	시간
Dialog	(1)	CO=Computer? * memory	49.8 초	PY=199?	2 분 28.7 초
	(2)	Computer? * memory	2 분 17 초	199?	1 분 43.7 초
Kipris	(3)	KW=컴퓨터? * 메모리	13.8 초	AD=1990~1998	1 분 35.2 초
		(3) - (1)	-36 초	(3) - (1)	-53.5 초
		(3) - (2)	-132.2 초	(3) - (2)	-8.5 초

* CO: Company name, PY=Publication Year, KW=키워드, AD=출원일자



단순비교에 따른 위험이 있기는 하지만, 비슷한 검색항목에 대한 응답속도는 Dialog 시스템에 비해 본 인터넷 시스템이 필드제한 검색에서는 평균 44.8초, 풀텍스트인 경우는 무려 1분 10초 가량 빠른 것으로 드러났다.

Interoperability with WWW/Database/Search Engine



kdl

Conclusion

- 기대효과
  - 연구개발 관련 중복투자 방지 및 신기술개발촉진을 통한 산업계의 기술경쟁력 강화
  - 특허제도 및 전자출원 등의 대민서비스 개선
  - SGML의 표준화로 데이터관리와 처리에 대한 비용감소
- 시스템개발
  - Prototype 개발경험을 통한 검색시스템의 신속한 개발
  - 관리기법/1(Method/1) 방법론 에 의한 프로젝트 진행
  - 국제표준 문서포맷인 SGML기반의 검색시스템과 SGML Viewing System 개발
  - 검색엔진과 데이터베이스의 연동과 분산환경의 시스템구축

kdl