Intelligent Search Engine

발표자 : 공기중(에이드텍 대표이사)

AidTech

Contents

AidTech 회 사 소 개

Product 개 요

AidSearch 시 스 템

AidSearch 기 능

AidRobot

AidDBI

AidFilter

AidCategorizer

AidSearch 활용시 장점

Reference Site

AidTech

회 사 소 개

1 page

회사개요

- 회사명: ㈜에이드텍
- 대표이사 : 공기중
- 자본금: 4억원
- 설립일: 2000년 1월 26일
- 임직원: 12명
- 사업분야: 인터넷/인트라넷 정보검색
  - Portal, EC, Contents, KMS
- 소재지 : 서울 서초 잠원동 37-12 논현 빌딩 501호
- 연락처 : Tel. 02-3445-4316 Fax : 02-3445-4318
  www.aidtech.co.kr

AT AidTech

회 사 소 개

2 page

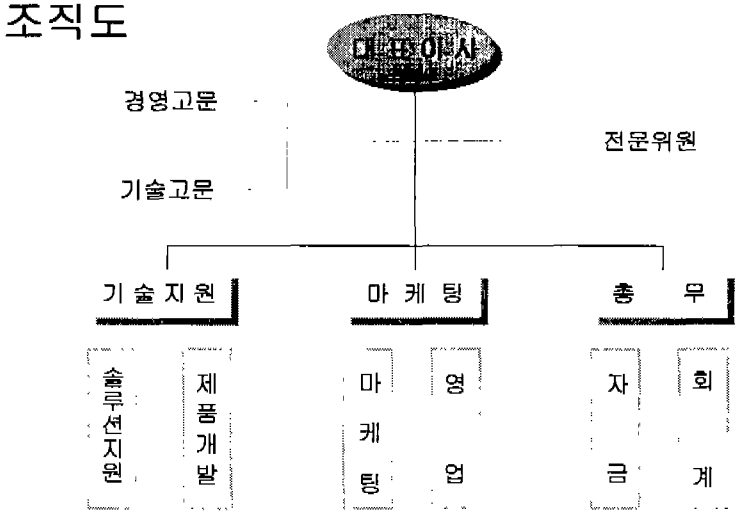
회사연혁

- ㈜ 에이드텍 설립 2000/01
- 대한매일 뉴스넷 뉴스검색계약 2000/03
- 옥션 상품검색 솔루션 제공 2000/04
- 삼성전자와 MOU 2000/05
- 스포츠서울 뉴스검색 구축 2000/06
- 필리핀의 유력 일간지 PhilStar Daily와 MOU체결 2000/06
- 삼성솔루션페어 2000 참여 2000/06
- 한겨레신문 뉴스검색 계약 2000/07
- B2B검색포털 바이어스타트 검색솔루션 제공 계약 2000/08
- POSCO 홈페이지 검색시스템 구축 계약(영어,한글) 2000/10
- 대법원 홈페이지 검색시스템 구축 계약(LG-EDS) 2000/11
- 국군 정보사령부 인터넷 군사 정보 몽향 검색시스템 계약 2000/11
- 서울시청 4개국어 문화관광포탈사이트 계약 2000/12
- 산업기능요원 지정업체 선정 2000/12
- 벤처기업 인증 2001/01

AT AidTech

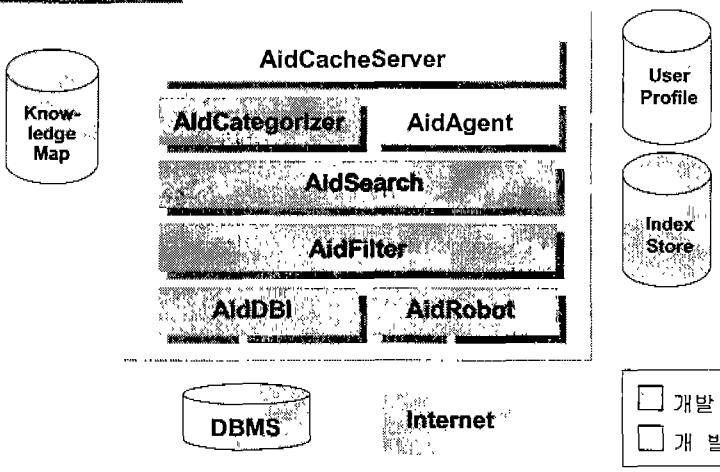
회 사 소 개

조직도



Product 개요

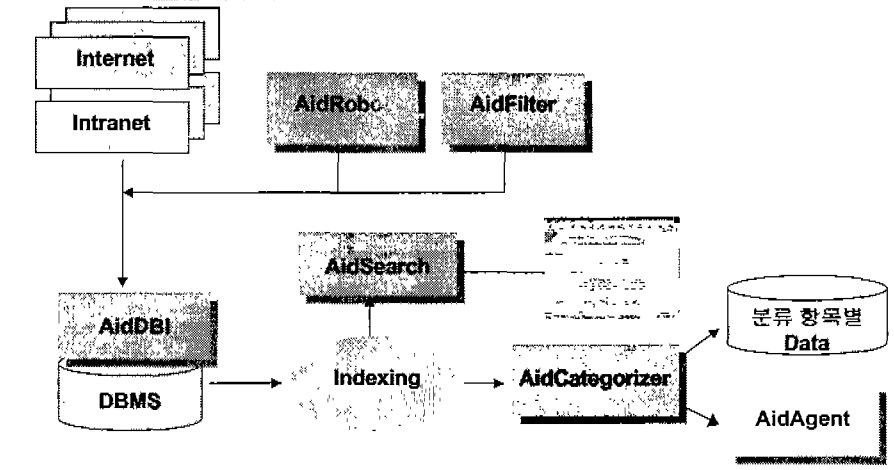
Product 구성도



Product 개요

5 page

Product 기능도



Product 개요

6 page

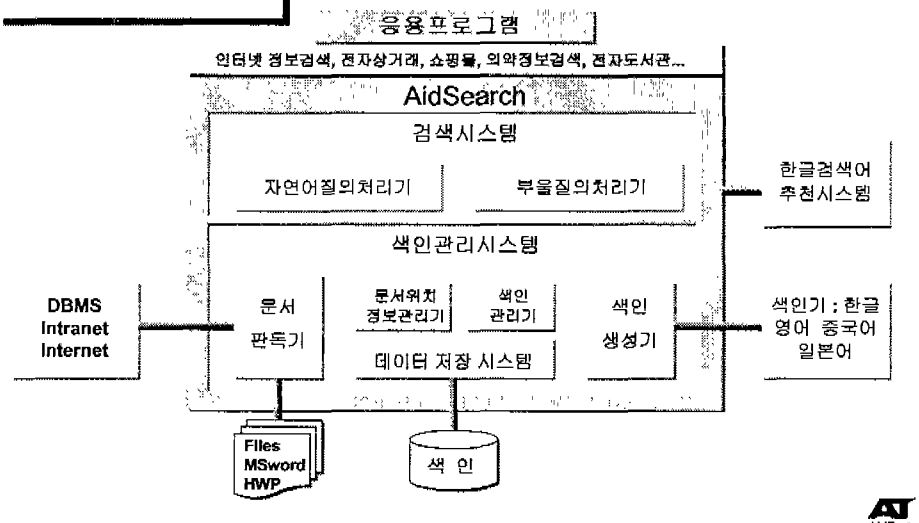
Product 기능

제 품 명	기 능
AidRobot	• 인터넷 URL 상의 Data를 수집
AidDBI	• DBMS Interface Tool(DBMS Reader)
AidFilter	• MS Office, HWP, Arirang등의 파일을 text파일로 변환
AidSearch	• Document에서 추출된 색인을 바탕으로 검색
AidCategorizer	• KnowledgeMap에 기반한 자동 분류
AidAgent	• Mail, 폴더로 자동배포 ※개발중인 제품
AidCacheServer	• 검색결과 Caching으로 동일 질의에 대한 검색 성능 향상 ※개발중인 제품

AidSearch 시스템

7 page

AidSearch 구성도



AidSearch 시스템

8 page

■ 검색 시스템

- 색인정보 및 통계 정보를 이용
- 사용자의 질의를 만족하는 문서를 검색

■ 데이터 저장 시스템

- 부울 질의 및 자연어 질의에 의한 검색을 지원
- 디스크와 같은 물리적 저장 장치에 대용량의 데이터를 저장할 수 있도록 도와 주는 저장 시스템
- 객체 파일, 인덱스 파일, 바이트 파일을 지원

AidSearch 시스템

■ 색인 관리 시스템

- 색인관리기 :
  문서로부터 추출된 색인어들을 역파일구조로 저장
  관리 색인어 통계정보, 동의어 파일, 불용어 파일들을 관리
- 토큰 생성기 라이브러리 :
  색인어들을 추출하기 위한 다양한 자동 색인 모듈들을 관리
- 문서 위치 정보 관리기 :
  파일 시스템, **DBMS**, 인트라넷, 인터넷 등에 존재
  하는 문서들의 위치 정보를 관리
- 문서 판독기 라이브러리 :
  문서 위치 정보를 이용하여 문서로부터 텍스트를
  추출하는 문서 판독기들을 관리

AidTech

AidSearch 시스템

AidSearch 특/장점

- 빠른 검색 및 색인 속도 ▶ 검색 포탈에 강점
- 대용량 데이터 신속 처리 ▶ **BtoB, Portal Site, EC, Contents Service**
- 다국어 지원 ▶ 한국어, 영어, 일본어, 중국어(年內 독어,불어,스페인어)
- 자연어 질의에 의한 검색 ▶ 생활어로 직접 검색 가능(**keyword**방식 기본)
- 다양한 검색 방법 지원 ▶ 절단검색, 제한검색,부울연산,유사문서검색 등
- 다양한 **H/W Platform**지원 ▶ **Unix, Linux, Windows N/T, AS/400**등
- 타 정보시스템과의 연동 ▶ **DBMS, EDMS, Groupware**등
- 공유메모리 기능을 활용한 빠른 색인 ▶ ***200MB/hr***
- 원천 소스 보유로 제품의 개발이 자유로움 ▶ 해외진출 가능

AidTech

AidSearch 시스템

AidSearch 인덱스 속도

- 데이터 : **Empas Data**
- 원본 데이터 사이즈 : 약 **12G**
- 색인 **DB** 사이즈 : 약 **12G**
- **CPU : Intel Pentium II 450MHz**
- 색인 시간: 약 **60**시간(시간당 **200M**처리)

AidSearch 시스템

AidSearch 검색 속도

■ 인텔칩(원본 데이터 : **12GB** 엠파스 데이터)

CPU: Intel Pentium II 450MHz **OS: Linux 2.2.12**
Memory: 1GB **HDD: SCSI IBM DRHS36V(7200RPM)**

전체질의건수 : **100000** 건
평균검색시간 : **762.40509 ms** (검색건수 0건 제외)
최대검색시간 : **4100.00000 ms**
검색건수 0건 : **8852** 건

초단위	시간대별 구간	질의 빈도수	비율	누적비율
[1]	1 - 1000 ms	63876	0.70	0.70
[2]	1001 - 2000 ms	25850	0.28	0.98
[3]	2001 - 3000 ms	1346	0.01	1.00
[4]	3001 - 4000 ms	74	0.00	1.00
[5]	4001 - 5000 ms	2	0.00	1.00

AidSearch 시스템

■ 알파칩

CPU: Compaq Alpha DS10 466MHz　　OS: Linux 2.2.13
Memory: 1GB　　HDD: SCSI IBM DPSS-336950N

전체질의건수 : 100000 건
평균검색시간 : 455.96014 ms (검색건수 0건 제외)
최대검색시간 : 2800.00000 ms
검색건수 0건 : 8852 건

초단위	시간대별 구간	질의 빈도수	비율	누적비율
[1]	1 - 1000 ms	88137	0.97	0.97
[2]	1001 - 2000 ms	2937	0.03	1.00
[3]	2001 - 3000 ms	74	0.00	1.00

- Intel Pentium III 600 MHz로 처리하면 Intel Pentium II 450보다는 빠르고 Alpha chip과 유사할 것으로 추정됨.
- 위의 검색속도 시간은 검색 후 검색 결과를 보여 주기까지 걸린 시간임.

AidTech

AidSearch 기 능

■부울 질의에 의한 검색

- 부울 연산(AND, OR, NOT)
- 근접도 연산(WITHIN, NEAR, PHRASE)
- 절단 연산(좌, 우, 중간, 양측)
- P-Norm 모델 기반의 질의-문서 유사도 계산을 통한 문서 순위 결정

구 분	종 류	표 기 법	기 능
부 울 연 산	AND	A&B	A와 B가 동시에 출현하는 문서들 검색
	OR	A|B	A와 B중 어느 하나라도 출현하는 문서들 검색
	NOT	A!B	A는 출현하고 B는 출현하지 않는 문서들 검색
근접도 연 산	WITHIN	A ~n B	A가 B에 선행하여 동일 문서내에 출현하며, 문서내에서의 A와 B사이의 상대적인 단어 거리가 정수값 n이내인 문서들 검색
	NEAR	A ^n B	A와 B가 순서에 상관 없이 동일 문서내에 출현하며, A와 B 사이의 상대적인 단어 거리가 정수값 n이내인 문서들 검색
	PHRASE	A B C	A,B,C가 하나의 phrase로서 인접하여 출현하는 문서들 검색
절 단 연 산	좌측	AB*	*: 임의 개수의 임의 문자와 일치하는 단어를 포함한 문서들의 검색
	우측	*AB	
	중간	A*B	?: 하나의 임의 문자와 일치하는 단어들를 포함한 문서들의 검색
	양측	*AB*	

AidTech

AidSearch 기 능

15 page

■ 자연어 질의에 의한 검색

- 최신 정보 검색 모델 기반의 질의-문서 유사도 계산을 통한 문서 순위 결정
- 최신 정보 검색 모델 기반의 문서-문서 유사도 계산을 통한 유사 문서 검색
- 적합성 정보를 활용한 질의 재구성
- 문서에 나타난 단어나 문구 등을 그대로 검색에 사용
- 검색어에 *, ?의 절단 연산자(좌, 우, 중간, 양측) 사용 가능

AidTech

AidSearch 기 능

16 page

■ 다양한 한글 자동 색인

- 어절 단위 색인
  - 불용어를 제외한 어절 또는 단어를 색인어로 선정하는 방식으로, 원문에 나타난 그대로를 색인어로 사용
  - '저자명'이나 '키워드'와 같은 필드의 색인 방식으로 유용
- 형태소 단위 색인
  - 한글 형태소 분석기를 이용하여 한글 문서의 각 어절에 대해 형태소 분석을 수행함으로써 명사, 조사, 접미사, 동사, 형용사 등의 최소 형태소 단위를 구분한 후, 한글 문서에서 중요한 의미를 갖는 단순 명사(simple noun)를 색인어로 추출
- N-Gram 기반 색인
  - 어절 단위 색인을 통하여 선정된 색인어들을 연속된 n개의 문자들로 분할하여 색인하는 방식으로, 고가의 형태소 사전을 사용하지 않으면서도 형태소 단위 색인과 유사한 검색 효과를 지원
- 색인 방식 커스터마이즈의 용이함
  - 기타 개발자가 제공하는 자동 색인 시스템들을 C 언어 수준에서 결합할 수 있음

AidTech

AidSearch 기 능

17 page

■ 다른 정보 시스템과의 연동

다음과 같이 다양한 정보 시스템에 다양한 형식으로 존재하는 문서들에 대한 문서 판독기를 플러그인형으로써 문서들을 접근하고, 문서에 대한 색인을 구축하여 텍스트 검색을 지원 가능

정보시스템	데이터 포맷
- DBMS - EDMS - 인트라넷 정보 시스템 - 인터넷 정보 시스템 - 그룹웨어 시스템 - XML 문서 관리시스템 등	Text HWP MS WORD : :

AidTech

AidSearch 기 능

18 page

■ 기타 주요 기능

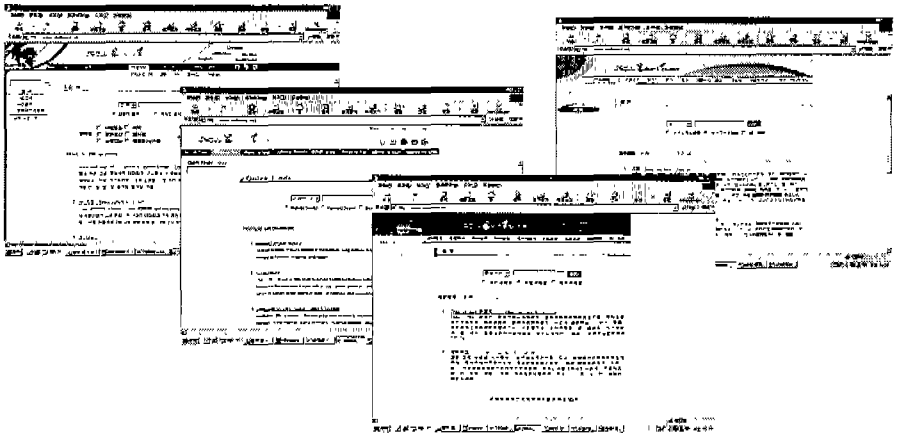
- TEXT, HWP, HTML, XML 등의 이종 구조 문서에 대한 검색
- 역화일을 이용한 빠른 색인어 접근
- 주기억장치를 활용한 빠른 인덱싱
- 이음동의어 검색
- 날짜 제한 검색
- 분야 제한 검색
- 검색 결과내에서의 재검색
- 질의 작성 지원을 위한 색인어 조회
- 질의 확장을 위한 검색어 추천 기능
- 전문 검색
- 전문의 문단 검색
- 한자/한글 변환 검색
- 대.소문자 구별 검색
- 날짜, 수치 등의 정형화 필드 검색
- 적합성 피드백에 의한 질의 수정
- 온라인 문서 삽입, 삭제
- 검색어 하일라이팅
- 검색된 문서에서의 질의 최적문단 추출
- 다양한 유닉스 플랫폼에서 사용 가능

AidTech

AidSearch 기 능

19 page

AidSearch Demo (4개국어)



AidRobot

20 page

AidRobot

WWW

웹 전체 검색시

Ex) 옥션, 네이버,

…

지정 URL

관련사이트 검색시

Ex) 의료포탈, 자동차포탈,

…

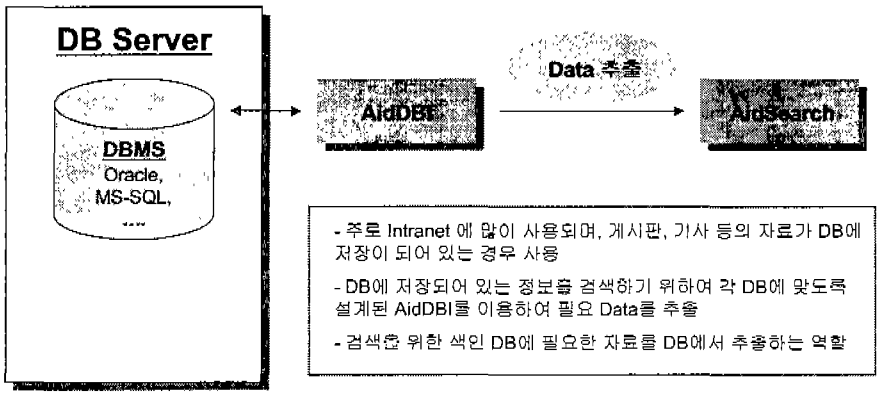
AidRobot → (HTML, HWP 등의 문서수집) → AidSearch

- Internet 에 사용되며, 웹에 올려져 있는 문서들을 가져와 색인 Data로 활용
- 웹 검색의 성격에 맞게 URL을 지정하여 검색 대상 Data를 수집 가능
- File에 대한 주소 정보 저장

AidDBI

21 page

AidDBI

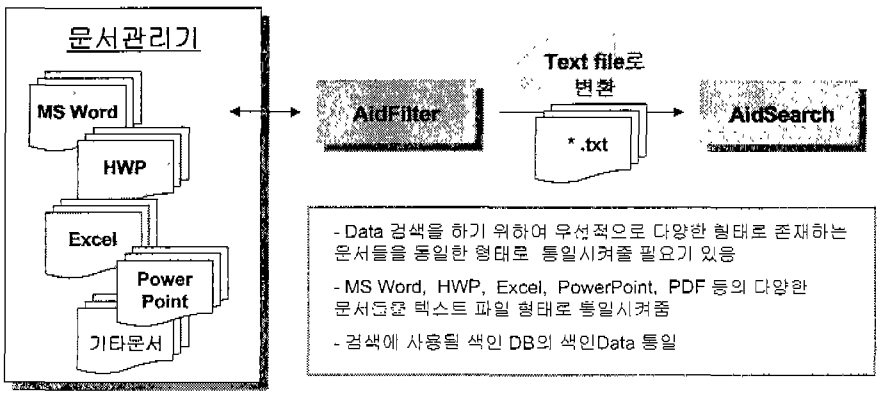


- 주로 Intranet 에 많이 사용되며, 게시판, 기사 등의 자료가 DB에 저장이 되어 있는 경우 사용
- DB에 저장되어 있는 정보를 검색하기 위하여 각 DB에 맞도록 설계된 AidDBI를 이용하여 필요 Data를 추출
- 검색을 위한 색인 DB에 필요한 자료를 DB에서 추출하는 역할

AidFilter

22 page

AidFilter



- Data 검색을 하기 위하여 우선적으로 다양한 형태로 존재하는 문서들을 동일한 형태로 통일시켜줄 필요가 있음
- MS Word, HWP, Excel, PowerPoint, PDF 등의 다양한 문서들을 텍스트 파일 형태로 통일시켜줌
- 검색에 사용될 색인 DB의 색인Data 통일

AidCategorizer

AidCategorizer 개요

■ 분류방식

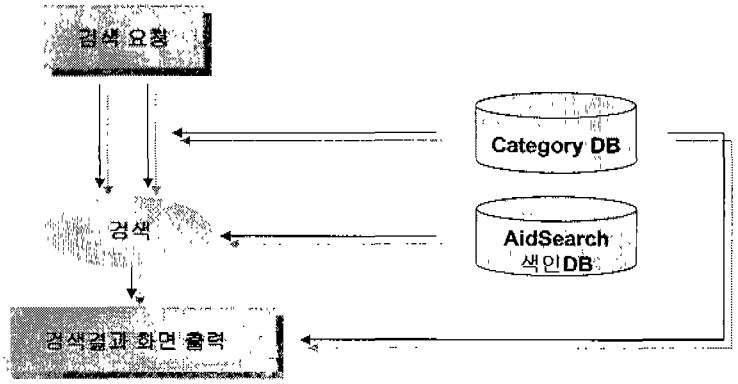
자동 분류의 방식은 일반적으로 분류기의 학습 방식과 Query-Base 방식이 있다.
AidCategorizer 는 AidSearch 색인 DB에 바탕을 둔 Query-Base 방식을 채택하고 있다.
Query-Base 방식이란 각 Category 에 해당하는 문서를 매핑 시킬때 그 기준을 Query에 두는 방식이다.

■ 구조

AidCategorizer 는 AidSearch가 설치된 시스템에서만 사용할 수 있다.
AidCategorizer 의 분류 DB도 AidSearch 의 색인 DB가 있어야 만들어 질 수 있다.
AidCategorizer 는 분류 DB를 만들 때 지정된 색인 DB에서 각 분류의 Query를 가지고 검색하여 각 분류에 어떠한 문서들이 해당하는지 결정한다. 프로그래머는 AidCategorizer 의 API를 이용하여 각 분류에 해당하는 문서들의 리스트를 얻어 올 수 있으며,
이 리스트를 가지고 색인 DB에서 제한 검색을 함으로서 분류 제한 검색을 행할 수 있다.

AidCategorizer

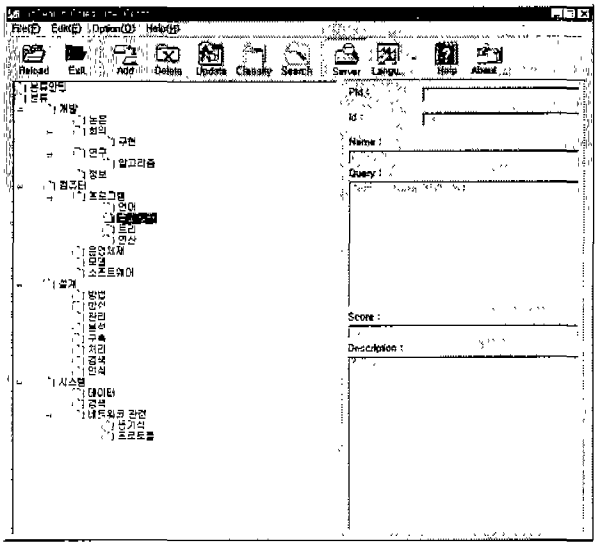
AidCategorizer 구조도



AidCategorizer

25 page

AidCategorizer Demo



AidProducts 활용시 장점

26 page

포탈사이트, 전자상거래, 언론사, 정부기관, BtoB, 커뮤니티, KMS, EDMS

Powered by AidSearch — 최고의 검색 솔루션 지원 — 검색

최고의 검색 기능	최고의 ReferenceSite	순수 국내 기술
• 빠른 검색 /색인속도 • 대용량 Data 신속처리 • 다국어 지원 • 자연어 지원 • 다양한 H/W지원 • Ranking Quality	• 검색Portal: empas, NAVER.COM • 전자상거래 : Auction • B2B포탈 : enstart • 언론기관 : 스포츠서울, Kdaily.com • PtoP사이트 : DBDIC.com	• 원천 소스 보유로 자유로운 개발 가능 • 한글 처리 강점 • 정부구매 사업참여 유리 • 다양한 검색 시스템 요구사항 수용 가능

Reference Site

- ***empas*** : 인터넷 검색 포털 사이트
- **NAVER** : 인터넷 검색 포털 사이트
- **Hirit.com** : 필리핀 포털 사이트

- 대한매일 뉴스넷 : 신문 기사 검색
- 스포츠서울 : 신문 기사 검색
- 한겨레신문 : 신문 기사 검색

- ***Auction*** : 상품 검색
- **BuyerStart** : B2B포털 사이트
- **DBDIC** : P2P사이트

- 대법원 : 홈페이지 검색
- 서울시청 : 문화광광과 4개국어 검색
  (한글, 영어, 일본어, 중국어)
- **POSCO** : 홈페이지 검색

■ 개발중

즐거운학교 : 사이트 검색

국군정보사령부 : 검색시스템 구축

MBC 뉴스 동영상 검색