Natural Language Processing Technology

고속, 고성능의 한국어 질의 응답 시스템

다이퀘스트닷컴

Intelligent web dialogue interface solutions

°김학수*, 이근배**, 서정연***

*㈜다이퀘스트 연구소
** 포항공대 컴퓨터공학과 자연어처리 연구실
*** 서강대학교 컴퓨터학과 자연어처리 연구실

diquest

Contents

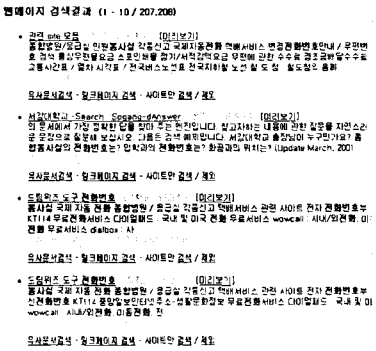
- Introduction
- Indexing method
  - System architecture
  - Answer-finding
  - Calculation of scores
- Searching method
- Experiments
- Demo
- Conclusion

Introduction (1/3)

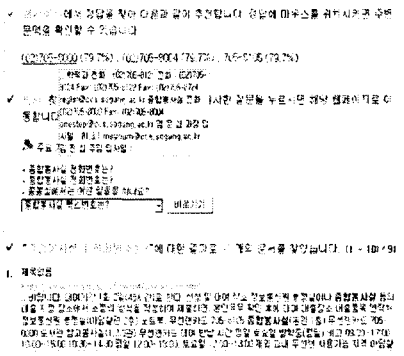
- **Traditional information retrieval system**
  - It returns a large amount of documents.
  - Users should carefully look over the whole text.
- **Question-answering system**
  - It searches a large collection of texts.
  - It filters out inadequate phrases or sentences.

Introduction (2/3)

- 종합봉사실 전화번호는?



IR system



QA system

Introduction (3/3)

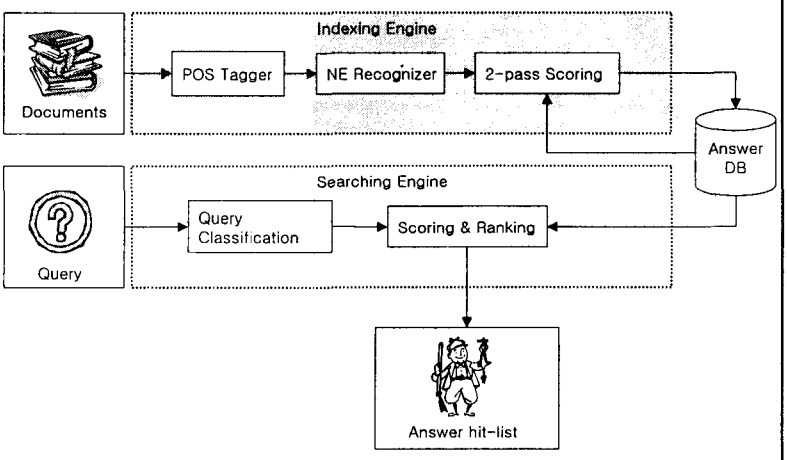
- Representative QA systems

	Strengths	Weaknesses
MURAX	Shallow linguistic analysis (POS tagger, finite-state recognizer)	Handcrafted patterns (Complex and fragile)
GuruQA	Fast response (Predictive annotation)	Use only local information
FALCON	High performance	Domain-specific knowledge like a semantic net.

1. 검색 속도가 IR 시스템에 비해서 상대적으로 많이 느리다.
   → 색인 필요
2. Passage 내의 정답 주변 정보만을 이용한다.
   → Second level 정보 이용 필요

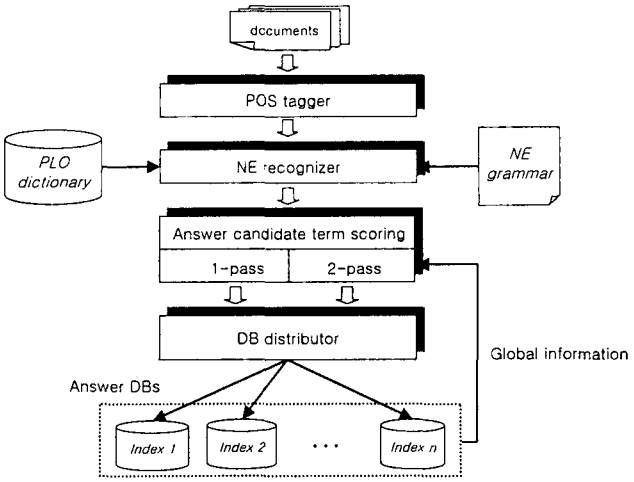
diquest

System architecture (1/2)



diquest

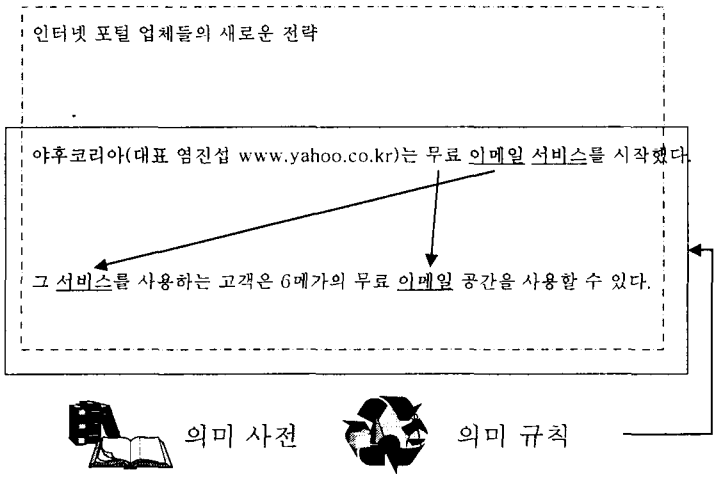
System architecture (2/2)



diquest

Answer-finding (1/2)

- 문맥 윈도 설정



diquest

Answer-finding (2/2)

- 105개의 의미 분류

계층 1	계층 2				
animal	bird	fish	mammal	person	reptile
plant	grass	tree	flower		
location	address	building	city	continent	country
	state	town			
identification	code	e_mail	nationality	position	tel_num
	URL				
date	day	month	season	weekday	year
time	hour	minute	second		
quantity	age	distance	duration	length	money
	number	power	rate	size	speed
	temperature	volume	weight		

Calculation of scores (1/3)

야후코리아(사장 염진섭 www.yahoo.co.kr)는 무료 이메일 서비스를 시작했다.
그 서비스를 사용하는 고객은 6메가의 무료 이메일 공간을 사용할 수 있다.

→ Distance, Term Frequency, TF*IDF

- 거리 가중치 (distance feature 반영)

$$distw_{d,k}(a_i, w_j) = \frac{c}{\log(dist(i,j)) + c}$$

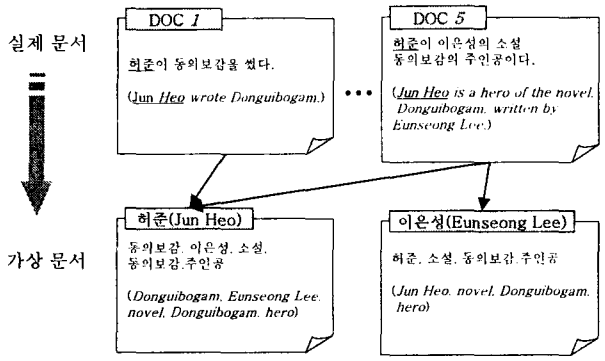
- 빈도수 (term frequency feature 반영)

$$LS^n_{d,k}(a_i, w_{pos(n)}) = distw_{d,k}(a_i, w_{pos(n)}) + \left(1 - distw_{d,k}(a_i, w_{pos(n)})\right) \times LS^{n-1}_{d,k}(a_i, w_{pos(n-1)}),$$
$$where\ LS^0_{d,k}(a_i, w_{pos(0)}) = 0.$$

Calculation of scores (2/3)

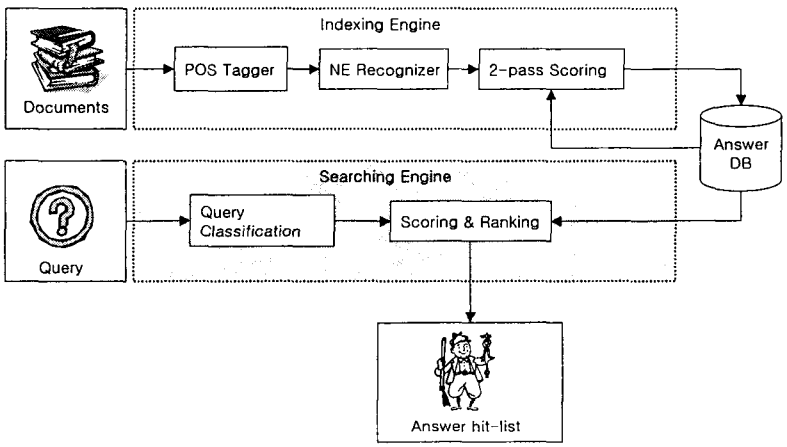
- **TF*IDF**

$$GS(pseudo\_d_a, w) = \begin{cases} \left(0.5 + 0.5\frac{tf_w}{Max\_tf}\right)\frac{\log(N/n)}{\log(N)}, & if \quad tf_w > 0 \\ 0 & , \ if \quad tf_w = 0 \end{cases}$$



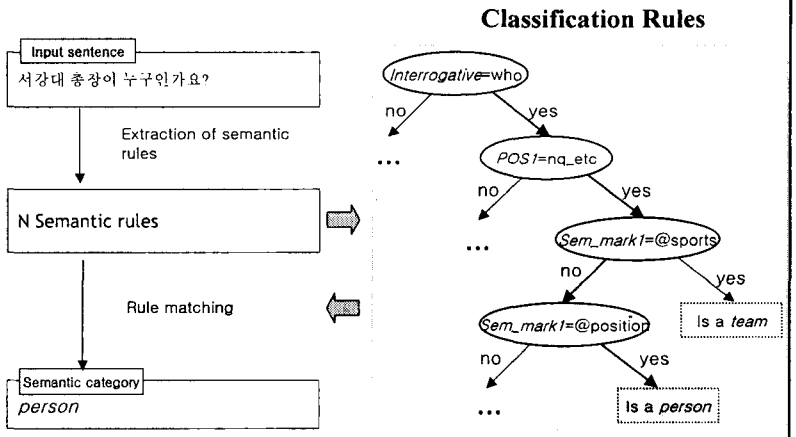
Intelligent web dialogue interface solutions

System architecture



diquest

Searching method (1/2)



Searching method (2/2)

- 정답 후보 점수 계산: p-Norm 모델 (AND)

$$Sim(A, Q_{and}) = 1 - \sqrt[p]{\frac{q_1^p(1-S(A,w_1))^p + q_2^p(1-S(A,w_2))^p + \otimes + q_i^p(1-S(A,w_i))^p}{q_1^p + q_2^p + \otimes + q_i^p}}$$

- IR, QA 유사도 결합

$$Sim(D,Q) = \frac{\alpha \cdot IRsim(D,Q) + \beta \cdot QAsim_d(A_{max}, Q)}{\alpha + \beta}$$

Experiments (1/2)

- 실험 데이터
  - KorQATeC 1.0
    - 207,067 balanced documents (368,768 kilobytes)
  - WEBTEC 1.0 (WEB TEst Collection)
    - 22,448 documents (43,953 kilobytes)
    - www.sogang.ac.kr (7,869 documents)
    - korea.internet.com (6,452 documents)

diquest

Experiments (2/2)

- 질의 응답 정확률 실험

	Lee2000 (object)	Lee2000 (50-byte)	Kim2001 (object)	실험시스템
MRAR	0.322	0.456	0.485	0.540
MRAR-1	0.322	0.456	0.539	0.600

- 질의 응답 및 색인 속도 실험

	Response time per query (seconds)	Indexing time per mega byte (seconds)
기본 IR 시스템	0.026	26.765
실험 시스템	0.048	30.542

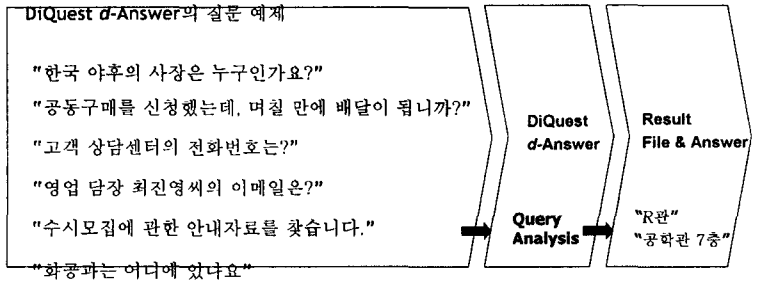
diquest

DiQuest *d*-Answer Enterprise : 지능형 웹 정보검색 에이전트

- DiQuest *d*-Answer의 특징
  - 자연어 검색 엔진인 DiQuest IR의 모든 기능에다 고객의 질문에 대한 정답을 추출하여 제공하는 지능이 추가된 자세대 검색 솔루션
  - 고객의 질문에 유사한 문서를 제시하는 일반적인 검색엔진과는 달리 DiQuest d-Answer는 문서 중에서도 질문에 대답이 되는 문장이 있는 문서만을 추려내어 정답 문장을 제시하는 솔루션

- DiQuest *d*-Answer의 질문 예제

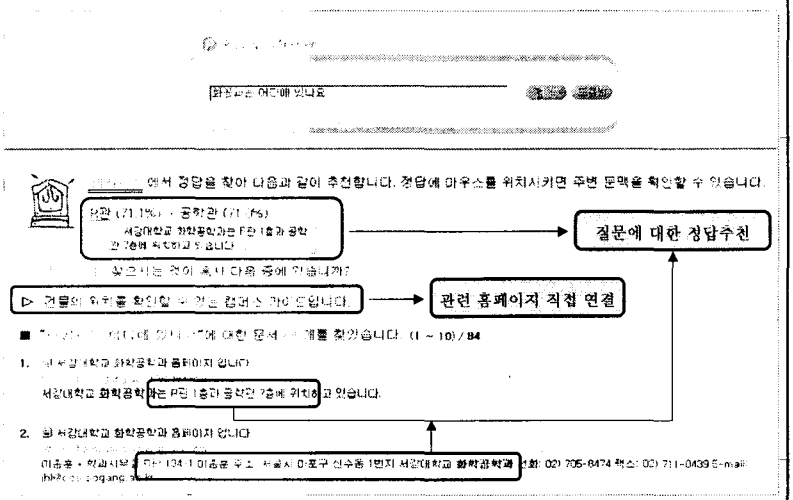
"한국 야후의 사장은 누구인가요?"
"공동구매를 신청했는데, 며칠 만에 배달이 됩니까?"
"고객 상담센터의 전화번호는?"
"영업 담장 최진영씨의 이메일은?"
"수시모집에 관한 안내자료를 찾습니다."
"화공과는 어디에 있나요"



diquest

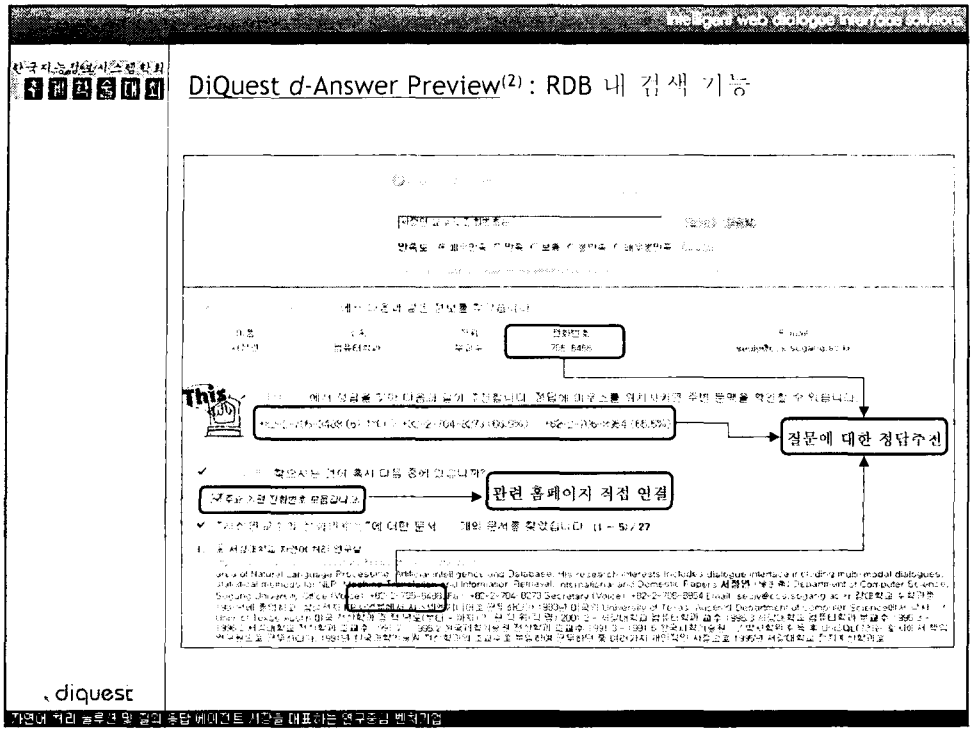
Intelligent web dialogue interface solutions

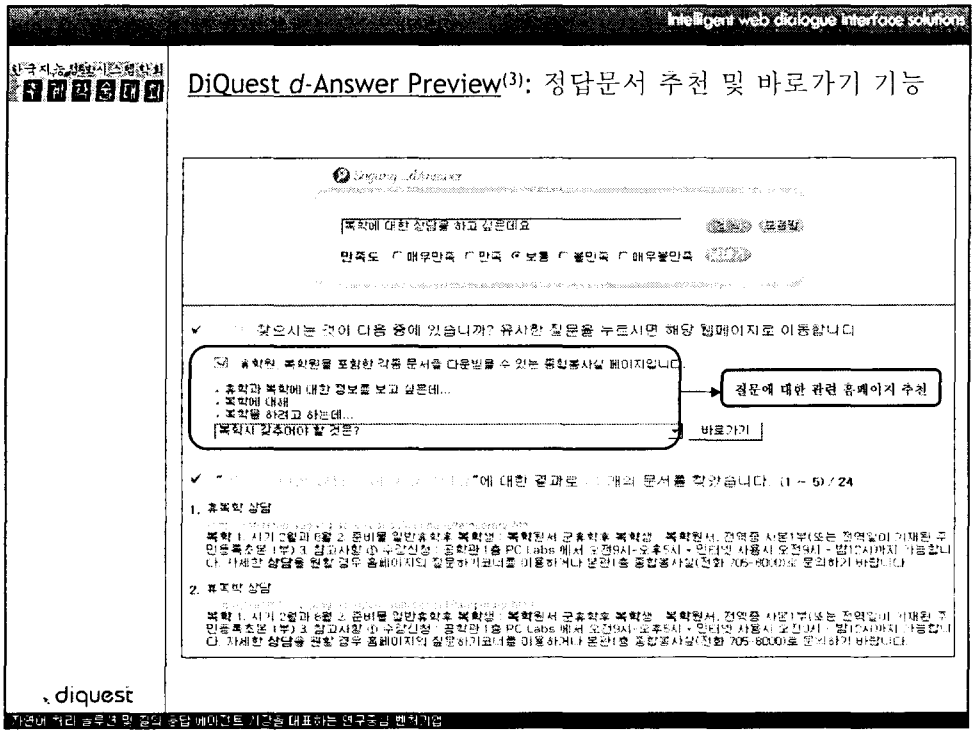
DiQuest *d*-Answer Preview(1) : 웹 문서상에서 검색

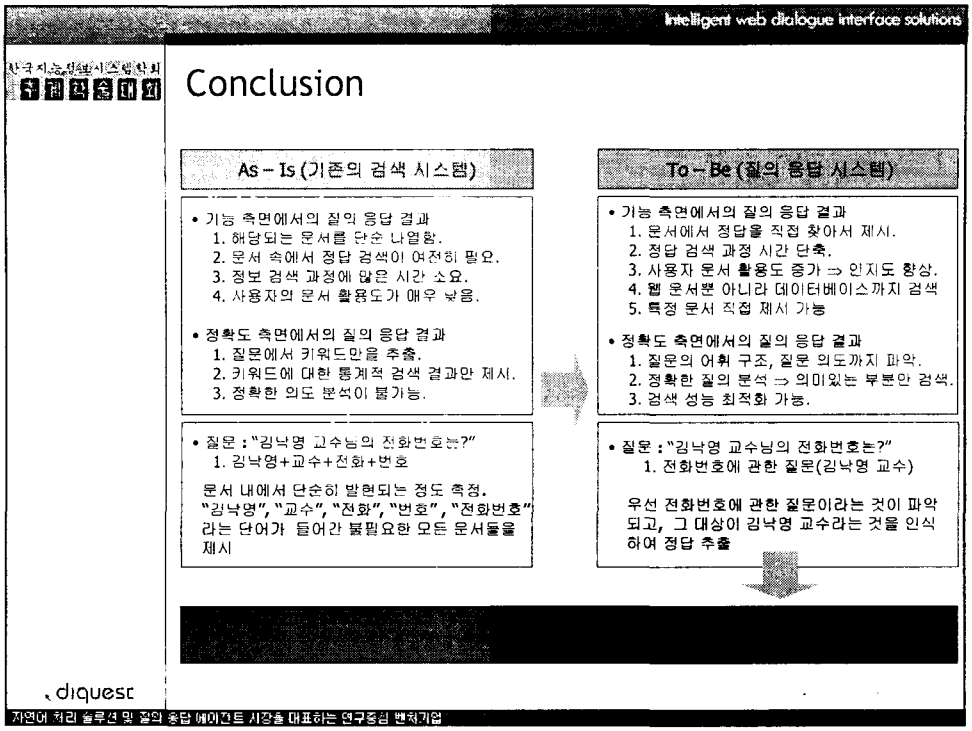


diquest

자연어 처리 솔루션 및 질의 응답 에이전트 시장을 대표하는 연구중심 벤처기업


Intelligent web dialogue interface solutions
DiQuest d-Answer Preview(2) : RDB 내 검색 기능
This
질문에 대한 정답추천
관련 홈페이지 직접 연결
diquest
가변어 처리 솔루션 및 질의 응답 에이전트 시장을 대표하는 연구중심 벤처기업


Intelligent web dialogue interface solutions
DiQuest d-Answer Preview(3): 정답문서 추천 및 바로가기 기능
찾으시는 것이 다음 중에 있습니까? 유사한 질문을 누르시면 해당 웹페이지로 이동합니다
질문에 대한 관련 홈페이지 추천
바로가기
diquest
가변어 처리 솔루션 및 질의 응답 에이전트 시장을 대표하는 연구중심 벤처기업


Intelligent web dialogue interface solutions
Conclusion
As – Is (기존의 검색 시스템)
• 기능 측면에서의 질의 응답 결과
1. 해당되는 문서를 단순 나열함.
2. 문서 속에서 정답 검색이 여전히 필요.
3. 정보 검색 과정에 많은 시간 소요.
4. 사용자의 문서 활용도가 매우 낮음.
• 정확도 측면에서의 질의 응답 결과
1. 질문에서 키워드만을 추출.
2. 키워드에 대한 통계적 검색 결과만 제시.
3. 정확한 의도 분석이 불가능.
• 질문 : "김낙명 교수님의 전화번호는?"
1. 김낙명+교수+전화+번호
문서 내에서 단순히 발현되는 정도 측정.
"김낙명", "교수", "전화", "번호", "전화번호"라는 단어가 들어간 불필요한 모든 문서들을 제시
To – Be (질의 응답 시스템)
• 기능 측면에서의 질의 응답 결과
1. 문서에서 정답을 직접 찾아서 제시.
2. 정답 검색 과정 시간 단축.
3. 사용자 문서 활용도 증가 ⇒ 인지도 향상.
4. 웹 문서뿐 아니라 데이터베이스까지 검색
5. 특정 문서 직접 제시 가능
• 정확도 측면에서의 질의 응답 결과
1. 질문의 어휘 구조, 질문 의도까지 파악.
2. 정확한 질의 분석 ⇒ 의미있는 부분만 검색.
3. 검색 성능 최적화 가능.
• 질문 : "김낙명 교수님의 전화번호는?"
1. 전화번호에 관한 질문(김낙명 교수)
우선 전화번호에 관한 질문이라는 것이 파악되고, 그 대상이 김낙명 교수라는 것을 인식하여 정답 추출
diquest
자연어 처리 솔루션 및 질의 응답 에이전트 시장을 대표하는 연구중심 벤처기업