□ 기술해설 □

# 에이전트 기반 정보검색†

포항공과대학교 이근배*·김동석·원형석·박미화

## 1. 서 론

소프트웨어 에이전트란 그 정확한 정의는 아직 내릴 단계는 아니지만 대략 사용자를 대신하여 유용한 일을 해주는 프로그램을 말한다. 과거의 객체기반(object-oriented) 소프트웨어의 개념에 자율성(autonomous), 목적 지향성(goal-directed), 협동성(coope-rative) 등의 새로운 개념을 추가하여 보다 능동적인 특성을 가지는 소프트웨어의 개발을 가능하게 해주는 소프트웨어 공학의 새로운 파라다임으로 볼 수 있다. 요즈음 이러한 소프트웨어 에이전트를 정보검색, 특히 인터넷 정보검색에 적용하려는 시도가 많이 일어나고 있다. 본고는 이러한 추세를 반영하여 에이전트 기반 정보검색의 필요성, 세계적인 연구현황 및 추세 그리고 포항공대에서 수행되고 있는 한국어 사용자를 위한 에이전트 기반 인터넷 정보검색 프로젝트인 AIR-Web(I, II)에 대하여 차례로 알아본다.

## 2. 에이전트 기반 정보검색의 필요성

일반적으로 정보검색은 여러 사이트에 분산되어 있는 서로 다른 형질의 데이타에 대해 원활한 검색을 수행하여야 한다. 특히 인터넷 환경에서 바람직한 정보검색 시스템은 데이타베이스, 그룹웨어, 멀티미디어처리 등의 다른 시스템 요소와의 통합 솔루션을 지원하여야 하며 검색뿐 만이 아니라 정보라우팅 및 여과, 브라우징 및 정보시각화, 정보추출(information extraction)등의 다양한 서비스를 아울러 제공해주어야 한다[1]. 이러한 분산 이질환경의 다양한 정보검색 서비스를 위해서는 소프트웨어 에이전트의 대표적 특성들인 적응성(adaptability), 분산성(distri- buted processing), 자율성(autonomy), 사회성(social ability) 등이 매우 중요한 역할을 할 수 있다. 특히 멀티 에이전트 모델은 정보검색을 위해 다양한 정보원으로의 병렬 분산 접근을 가능하게 해주며 특정 정보에 맞는 특화된 정보검색도 가능하게 해 준다. 또한 인터넷의 방대한 크기에 맞는 확장성을 자연스럽게 제공할 수 있기 때문에 정보검색에서 에이전트의 도입은 필수적이며 선진국에서는 이미 에이전트와 정보검색을 결합한 연구가 많이 수행되고 있는 실정이다[22].

## 3. 국외 연구사례 및 추세

미국을 비롯한 선진국에서는 에이전트에 기반한 정보검색 시스템이 각각 에이전트의 특성과 해당 서비스의 종류에 따라 여러 가지로 개발되고 있다. 특히 인터넷상의 정보를 자율적으로 혹은 사용자의 지시를 받아 검색해주는 소프트 로봇(software robot)들이 많이 연구되어 MetaCrawler[2], Savvy- Search[3] 등의 메타서치엔진(meta-search engine)과 Shopbot[4], BargainFinder[5], Netbot Jango[6]같은 전자상거래를 위한 샵로봇(shopbot) 등의 시스템들이 개발되어 있다. 에이전트의 협동성을 이용해 성향이 비슷한 사용자들을 이용하여 정보를 여과해주는 새로운 개념의

† 본 연구는 과기처 특정과제인 소프트 과학프로그램(96~99)에 의해 부분 지원을 받은 것임.

* 종신회원

협동여과(collaborative filtering) 방법론도 개발되어 FireFly[7] 같은 상업용 시스템과 서비스도 수행되고 있다. 에이전트의 목표지향성을 살려 능동적으로 사용자에게 정보를 가져다주는 푸시(push)기술도 개발되어 SIFT[8], Pointcast[9]같은 정보여과 시스템에 사용되고 있는 실정이다. 한편 에이전트의 특성인 확장성(scalability)를 살리기 위한 에이전트기반 정보검색 프레임워크(framework)에 대한 기초 연구도 대학과 연구소를 중심으로 많은 진전을 보고 있으며 여기에는 미시간 대학의 전자도서관 프로젝트인 UMDL(Univ. of Michigan Digital Library) 구조[10], MCC의 InfoSleuth[11] 그리고 메릴랜드 대학의 CARROT/CAFE[12] 등의 대표 연구가 있고 이들 모두 broker, back-end agent, user agent의 기본 구조를 제시하고 있다. 이들 에이전트 프레임워크 연구는 모두 ARPA의 지식공유노력(know-ledge sharing effort : KSE)[13]의 결과인 KQML(knowlege query manipulation language)와 KIF(knowledge interchange format)을 이용하고 있는데 이중 KQML은 이미 에이전트들의 통신을 위한 표준 상위레벨의 프로토콜로 자리를 잡아가고 있으며 KIF 역시 지식기반 프로그램들을 위한 공통의 지식표현 언어로서 사용되고 있는 추세이다. 이러한 KQML통신 구조를 제공하기 위한 라이브러리들도 속속 개발되고 있으며 점차 KATS(Loral/UMBC), KAPI(Lockheed/EIT/Stanford) 같은 C/C++ 언어기반의 패키지에서 Java Agent Template(Stanford)[14] 같은 java 언어기반의 KQML API(application program interface)를 제공하는 패키지들로 바뀌어 가고 있다.

## 4. 포항공대 모델

국내에서도 에이전트연구의 필요성이 인식되어 국가 프로젝트인 소프트과학 연구에서 에이전트관련 과제[15]를 수행하고 있으며 특히 멀티 에이전트 구조, 휴먼 인터페이스 에이전트 및 인터넷 정보검색을 위한 적응(adaptive) 에이전트 등의 연구가 활발히 수행되고 있다. 그외에 ETRI 등의 연구소에서 에이전트 국제표준 단체인 FIPA(Foundation for Intelligent Physical Agent)[16]의 활동에 참여하고 있으며 Stanford 대학과 공동으로 Black-Board 구조기반 멀티에이전트 환경을 개발한 바도 있다.

이중에 본고에서는 포항공대에서 개발중인 인터넷 정보검색을 위한 한국어 자연어 정보검색 에이전트 시스템인 AIR-Web I, AIR-Web II를 자세히 설명하고자 한다.

### 4.1 AIR-Web I

#### 4.1.1 AIR-Web I 구조

AIR-Web[1][17]은 포항공대 HCI 랩에서 개발해온 한국어 자연어 처리를 기반으로 하는 정보검색용 멀티에이전트 환경이다. 통상적인 웹 정보검색 모델과는 다르게 AIR-Web은 세션을 단위로 하는 자연어 질의에 대한 정보검색을 지원하는 웹 검색도구이다.

전체적인 시스템 구조는 그림 1과 같으며 에이전트간의 통신을 위하여 KQML[18]를 채택하였다. 이를 위한 기본적인 API로는 JATLite를 활용하여 개발하였다. JATLite는 모든 기능이 자바언어로 구현되어 있으며 KQML 메시지 통신 및 에이전트 템플릿을 제공한다.

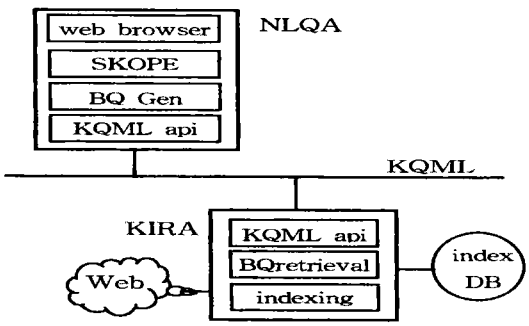


그림 1 AIR-Web I 전체구조

AIR-Web I에서는 기본적으로 NLQA(Natural Language Query Agent)와 KIRA(Korean Information Retrieval Agent)를 제공한다. NLQA는 netscape같은 웹 브라우저를 통하여 사용자의 자연어 질의를 받아서 부울린 질의어로 바꾸어 주는 역할을 하며 포

1) Agent-based Information Retrieval on the Web

항공대에서 개발된 한국어 해석엔진인 SKOPE (Standard KOrean Processing Engine)를 사용한다. SKOPE[19]는 한국어 자연어 입력에 대해 형태소분석, 품사태깅, 구문분석 및 의미분석을 하는 일반적인 한국어 언어분석 패키지로써 대화체 언어에 좋은 분석결과를 보여준다. NLQA는 SKOPE의 형태소 분석/품사태깅/구문분석 결과를 이용하여 자연어 질의를 해석하며, 결과로 나온 구문구조는 불리언 질의 생성기(BQ gen)에 의하여 불리언 질의로 바뀐후 KQML api를 통하여 에이전트 통신용 KQML 메시지로 변환되어 KIRA에게 보내진다. 한국어 정보검색 에이전트인 KIRA는 KQML message를 해석하여 불리언 질의를 뽑아내고 이를 이용하여 정보검색을 수행한다. 불리언 질의는 p-norm 모델을 통하여 index term과 정합되고 해당문서가 점수순으로 정렬되며 정렬된 문서는 KQML 메시지로 변환되어 NLQA의 웹 브라우저를 통하여 사용자에게 보여진다. KIRA의 색인(indexing)은 역시 SKOPE 엔진을 이용하여 명사 단일어 및 복합명사를 색인어로 추출하여 용어의 출현빈도수(term frequency : tf)와 역문헌 빈도(inverse document frequency : idf)를 이용하여 weighting되어 index DB에 저장된다.

#### 4.1.2 KIRA 색인(Indexing)

색인이란 문서에서 필요한 정보를 추출하여 검색시에 사용하도록 하는 과정이다. 입력 문서들의 구성 문장들을 입력으로 받아 형태소 분석 및 태깅 과정을 거치고 난 뒤 구문분석의 언어분석 단계를 거친다. 형태소 및 태깅과정에서 복합어 분할 및 단일어가 골라지고 구문분석 과정에서 구절(phrase)로부터의 복합어 합성이 이루어진다. 이렇게 선정된 색인어들 중에서 색인어로 가치가 없는 단어들을 불용어 처리한다. 남은 색인어들로 dictionary file 및 posting file을 구성한다. dictionary file은 색인어, 문서 출현빈도수(document frequency), 각 색인어에 대한 posting file의 offset을 가지고, posting file은 각 색인어에 대한 문서번호와 출현빈도수(term frequency)를 가지게 된다. 색인의 결과로 만들어지는 dictionary file과 posting file은 검색에 이용된다. 그림 2는 색인과정을 보이고 있다.

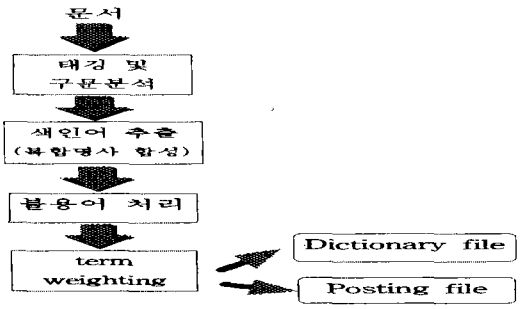


그림 2 색인 과정

#### 4.1.3 KIRA 검색(Retrieval)

검색은 자연어 질의를 기본 입력으로 하여 질의어의 구문분석 정보를 바탕으로 불리언 질의로 자동 변환하여 검색을 수행한다. 자연어 질의로부터 불리언 질의를 생성하는 이유는 일반 사용자로 하여금 자연어 질의를 가능하게 함으로써 사용의 편리를 도모하는 한편 사용자의 의도를 정확히 표현할 수 있는 불리언 질의로 검색 성능을 높이고 또한 여러 가지 검색모델에 적용이 가능한 장점이 있기 때문이다.

사용자로부터 입력된 자연어 질의는 범주문법에 기반한 구문분석 과정을 거친다. 구문분석 결과 생성된 구문 트리에서 검색어 및 연산자 정보를 추출한 후 복합명사 합성 및 term weighting과정을 거쳐 불리언 질의를 생성한다. 이렇게 생성된 불리언 질의어로 색인에서 생성된 dictionary file과 posting file을 이용해서 색인어와 검색어의 유사도 계산 후 랭킹한 검색결과를 사용자에게 보인다. 검색과정은

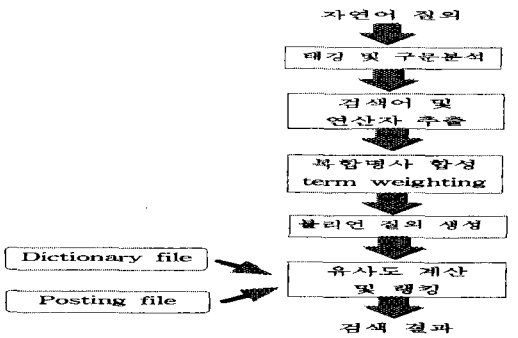


그림 3 검색 과정

그림 3과 같다.

## 4.2 AIR-Web II

### 4.2.1 AIR-Web II 구조

AIR-Web II는 다양한 이형질 분산 정보검색 에이전트를 포함하도록 AIR-Web I 멀티에이전트 구조의 확장된 형태를 갖는다. KQML의 통신 능력과 지식공유언어인 KIF[20]의 의미구조 응용을 위하여 NLQA에 한국어 의미/담화해석 기능을 추가하여 내용검색을 한국어로 할 수 있는 한국어 자연어 대화 에이전트(NLDA: natural language dialog agent)로 확장하였다. 특정 도메인에 대한 대화처리를 필요로 하는 응용 에이전트로 여러 도메인의 상품구매 도우미 에이전트(SAA: shopping aid agent)들을 추가하여 확장성을 보증할 수 있는 멀티에이전트 환경을 구축하였다. 즉, 두 개 이상의 에이전트가 상호 협동하여 목표를 해결하는 진정한 의미의 멀티에이전트 환경이 됨으로써 AIR-Web I에서는 필요 없었던 전체 에이전트의 정보를 관리해주는 조정자 에이전트(FA: facilitator agent)를 도입하였다. 전체 구조는 그림 4와 같다.

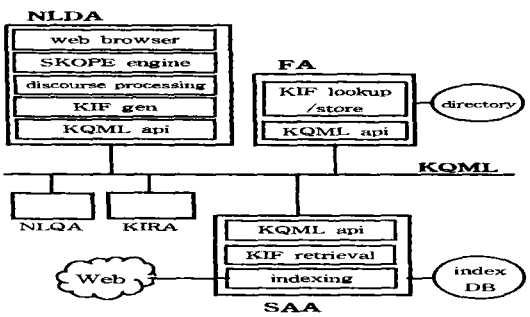


그림 4 AIR-Web II 전체 구조

### 4.2.2 자연어 대화 에이전트(NLDA)

NLDA는 NLQA에 추가로 의미/담화처리기능을 갖게 되며 조정자 에이전트와 함께 협동할 에이전트 정보를 받아서 사용자의 질의를 처리하게 된다. 담화분석기는 자연어 문장에서 문맥과 담화개체(discourse entity)를 고려하여 생략(ellipsis)과 대용현상(anaphora)를 처리하여 자연언어로 표현되어 있는 사용자의 의도를 충실히 파악할 수 있도록 해준다. 또 양화사(quantifier)의 스코핑(scoping)을 처리하여 문장의 애매성(ambiguity)을 제거한다. 이러한 생략, 대용, 스코핑의 해결은 사용자가 시스템과 대화하듯이 일련의 연관된 자연어 질의를 던질 수 있도록 해주는 세션기반의 자연어 질의를 지원한다[17].

사용자로부터 입력된 자연어 질의는 SKOPE 엔진의 의미분석과정을 거친 후 도메인 독립적인 담화분석을 통하여 생략/대용과 스코핑이 해결된 의미구조(semantic structure)인 유사 논리구조(QLF: quasi logical form)[21]로 출력된다. 담화분석까지 완료된 QLF 문장은 에이전트들에게 공통으로 사용되는 지식표현 언어인 KIF(know-ledge interchange format)로 번역된다. 이때 해당되는 도메인 지식을 이용하여 남아있는 애매성을 한번 더 걸러준다. KIF로 번역된 자연어 질의는 KQML 메시지에 embedding 되어 FA로부터 서비스가 가능하다고 확인된 SAA에 전달된다.

### 4.2.3 쇼핑 도우미 에이전트(SAA)

SAA는 특정 도메인에 대한 쇼핑 정보를 서비스해 주는 부분과 웹에서 제품 정보를 가져오는 웹 로봇부분으로 나뉜다. 웹 로봇은 특정 상품에 대한 정보를 가져와 HTML parser로 내용을 분석한 다음 필요한 정보를 패턴매칭 방식으로 추출하여 인덱스 DB에 저장한다. 상품 정보를 제공하는 인터넷 사이트별로 HTML 파일의 구성 형식이 다르기 때문에 인터넷 사이트별로 해당 정보를 해석할 수 있는 패턴을 도입한다. 이 파일을 패턴참조 파일이라고 하며 상품 구매 정보를 제공하는 사이트별로 참조 내용이 서로 다르다. 참조 내용은 모두 HTML 언어로 작성하여 정보 제공 사이트의 내용 변경에도 효과적으로 대응하여 상품 정보를 추출할 수 있도록 한다.

웹 로봇이 인덱스 DB에 저장되어 있는 정보를 업데이트시키는 시점은 두 가지 경우가 있는데 하나는 SAA 에이전트 사양에 명시된 업데이트 스케쥴링 정보에 의한 것이고 다른 하나는 외부 에이전트의 정보탐색 요청에 부합하는 정보를 보유하고 있지 않은 경우이다. 후자의 경우에는 직접 인터넷 사이트를 검색하여

요청된 검색을 수행하여 결과를 보내주고 아울러 인덱스 DB도 업데이트 하게 된다.

자연어 대화 에이전트가 특정 도메인에 관한 상품정보 검색을 요청했다면 SAA에게 해당 내용을 KIF 언어로 표현한 다음 KQML 메시지 형태로 wrapping하여 송신하게 된다. 이 메시지를 수신한 SAA는 우선 KQML 메시지를 파싱하여 KIF 내용을 뽑아낸 다음 다시 KIF 파서를 이용하여 전달 받은 내용을 분석하게 된다. 분석된 결과를 이용하여 인덱스 DB를 검색하기 위한 SQL(structured query language) 질의언어로 변환한다. 검색 결과는 상기의 역과정을 거쳐서 최종적으로 KQML 메시지 형태로 만들어진 다음 서비스를 요청한 NLDA에 전달된다.

#### 4.2.4 조정자 에이전트(FA)

AIR-Web I에서처럼 에이전트의 개수가 두 개인 경우에는 에이전트 상호간의 조정 역할이 필요 없을 수도 있다. 그러나 AIR-Web II는 두 개 이상의 본격적인 멀티에이전트 환경으로 설계되었기 때문에 에이전트 상호간의 조정 역할을 하는 에이전트의 필요성이 대두된다. 여러 개의 에이전트에 원활한 서비스를 제공하기 위하여 조정자 에이전트(FA)는 다음과 같은 대표적인 기능을 갖도록 설계한다.

- 에이전트의 등록 및 삭제
- 에이전트 브로커링(brokering)
- 에이전트 name service

모든 에이전트는 자신이 기동하면서 자신의 이름과 KIF로 표현된 서비스 내역을 FA의 디렉토리에 등록하게 된다. FA는 KIF를 해석하여 등록을 요청한 에이전트의 이름과 서비스 내용을 동적으로 데이타베이스에서 관리하게 된다. 그리고 해당 에이전트가 종료될 때 이름 및 서비스를 동적으로 삭제하게 된다.

FA는 에이전트 관련 정보를 관리하고 있기 때문에 그 내용을 기반으로 하여 정보를 제공하는 에이전트와 정보를 활용하는 에이전트간의 브로커 역할을 수행한다. 즉 상호간의 요청과 서비스에 대한 match making을 제공하는 것이다.

에이전트의 이름은 내부적으로는 IP주소와 포트 번호를 갖는 물리적 주소로 관리되지만 각각의 에이전트가 기동할 때 등록한 심볼릭 이름도 동일하게 관리한다. 따라서 white page에 해당하는 name service가 가능해지며 에이전트는 이것을 활용할 수 있다.

이외에도 FA는 에이전트 상호간의 지식표현 체계의 상이성 문제를 해결하기 위한 서비스로 다수의 vocabulary 사이의 변환 기능을 필요로 한다. 이러한 공통의 온톨로지(ontology) 서비스는 이형질 에이전트 시스템간의 상호작동성(inter-operability)을 보장하는데 있어서 매우 중요한 요소가 된다.

## 5. 결 론

본고에서는 정보검색에서 날로 중요성이 커져가는 소프트웨어 에이전트에 대하여 고찰하였다. 구체적으로 인터넷 정보검색에 왜 소프트웨어 에이전트가 도움이 되는가를 살펴보고 국제적인 연구동향과 포항공대에서 수행되고 있는 에이전트 기반 정보검색 시스템인 AIR-Web에 대하여 자세히 알아보았다. 앞으로 모든 소프트웨어는 에이전트화 될 것이며 인터넷을 환경으로 서로 협동하면서 사용자의 요구에 응하게 될 것이다. 이러한 환경에서 사용자와 에이전트간의 자연어를 이용한 의사전달도 핵심적인 기술요소가 된다. AIR-Web은 이 모든 기술을 개발할 수 있는 테스트베드로서 중요한 역할을 담당할 수 있다.

## 참고문헌

[1] W. B. Croft. What do people want from information retrieval?, D-Lib magazine, Nov. 1995(http://www.dlib.org/dlib/ november95/11croft.html).
[2] HREF http://www.metacrawler.com.
[3] HREF http://guaraldi.cs.colostate.edu: 2000/.
[4] HREF http://www.cs.washington.edu/ research/shopbot/.
[5] HREF http://bf.cstar.ac.com/bf/.
[6] HREF http://www.jango.com.

[7] HREF http://www.firefly.net.
[8] Tak W. Yan, Hector Garcia-Molina, SIFT-A Tool for Wide-Area Information Dissemination, Proceedings of the 1995 USENIX Technical Conference, pp. 177-186, 1995..
[9] HREF http://www.pointcast.com.
[10] E. H. Durfee, D. L. Kiskis, W.P. Birmingham, The agent architecture of the University of Michigan digital library, in M. Huhns and M. Singh(edited) Readings in agents, Morgan Kaufmann, 1998.
[11] R. Bayardo Jr. et. al., InfoSleuth: Agent-based semantic integration of information in open and dynamic environments, in in M. Huhns and M. Singh (edited) Readings in agents, Morgan Kaufmann, 1998.
[12] R. Scott Cost et. al. Agent development support for Tcl, internal research note, University of Marland Baltimore County.
[13] R. S. Patil et. al., The DARPA knowledge sharing effort: Progress report. In Principles of knowledge representation and reasoning: Proceedings of the third international conference(KR92), Morgan Kaufman, 1992.
[14] HREF http://java.stanford.edu.
[15] 박영택, 이근배, 최중민. 에이전트 속성 및 설계에 관한 연구. 과기처 국책 연구사업 (소프트과학) 1차년도 보고서, 1997.
[16] HREF http://drogo.cselt.stet.it/fipa/.
[17] Geunbae Lee, Jong-Hyeok Lee, Hyuncheol Rho. Natural language processing for session-based information retrieval interface on the web. Proceedings of IJCAI-97 workshop on AI in digital libraries, pp. 43-48, Nagoya, Japan, 1997.
[18] External Interface Working Group ARPA Knowledge Sharing Initiative, Specification of the KQML agent communication language, Working paper, Available as http://www.cs.umbc.edu/kqml /papers/kqmlspec.ps, Dec. 1992.
[19] 이원일, 김병창, 차정원, 이승우, 이근배, 이종혁. SKOPE: 한국어 음성언어 해석 엔진. 97 지능기술 expo, 한국 과학기술회관 대강당, 1997, 11월.
[20] M. R. Genesereth, R. E. Fikes, Knowledge Interchange Format Version 3.0 reference manual, Logic Group report Logic-92-1, Stanford Univ. Jun. 1992.
[21] Alshawi H., van Eijck J., Logical forms in the core language engine, In proceedings of the 27th Annual Meeting of the ACL., 1989.
[22] T. Finin, C. Nicholas, J. Mayfiled, Software agents for information retrieval, ACM SIG-IR97 tutorial notes, available at http://www.cs.umbc.edu/abir.

이 근 배

1984 서울대학교 컴퓨터공학과 졸업
1984~1986 서울대학교 연구조교
1986 서울대학교 컴퓨터공학과 석사
1987~1991 UCLA 전자계산과와 생명과학학과에서 연구조교와 연구원으로 근무
1991 미국 UCLA 전자계산학과 박사
1991~현재 포항공대 부교수
관심분야 : 자연어처리, 인공지능/신경망, 음성인식, 정보검색 등
E-mail:gblee@vision.postech.ac.kr

김 동 석

1986 경희대학교 전자공학과 졸업
1988 경희대학교 전자공학과 석사
1990~현재 LG전자 미디어통신 연구소 선임연구원
1998~현재 포항공대 대학원 전산과 재학
관심분야 : 에이전트시스템, 자연어처리, 정보검색
E-mail:dskim@pascal.postech.ac.kr

원 형 석

1997 경북대학교 컴퓨터공학과 졸업
1997~현재 포항공대 대학원 전산과 재학
관심분야 : 정보검색, 자연어처리
E-mail:moho@pascal.postech.ac.kr

박 미 화

1989 동아대학교 컴퓨터공학과 졸업
1989~1994 포스데이타 근무
1997~현재 포항공대 정보통신 대학원 재학
관심분야 : 정보검색, 자연어처리
E-mail:bfpark@pascal.postech.ac.kr

## ● '98 정보통신 하계워크샵 ●

- 일 자 : 1998년 8월 20일(목)~21일(금)
- 장 소 : 온양그랜드호텔
- 주 최 : 정보통신연구회
- 문 의 처 : 충남대학교 컴퓨터공학과 최 훈 교수
  Tel. 042-821-6652
  E-mail : hchoi@comeng.chungnam.ac.kr

## ● 제10회 한글 및 한국어 정보처리 학술대회 ●

- 일 자 : 1998년 10월 9일(금)~10일(토)
- 장 소 : 고려대학교
- 주 최 : 한국어정보처리연구회·한국인지과학회
- 문 의 처 : 서강대학교 전자계산학과 서정연 교수
  Tel. 02-704-8273
  홈페이지 : http://nlparies.sogang.ac.kr/~klip 98