감성용어 및 패턴을 이용한 감성기반 분산 문서분류시스템

김명규 ° 인주호 채수환
(kimmk°, allpoyou, chae@hau.ac.kr)
한국항공대학교 컴퓨터공학과

Distributed Document Classification System using Susceptibility Terms and Patterns

MyungKyu, Kim, ° JooHo In, SooHoan Chae
Dept Computer Engineering , Korea AeroSpace University

ABSTRACT

인터넷이 폭 넓게 보급되어 개인의 의견을 개진할 기회가 확대됨에 따라 정치,경제 등의 사안이나 제품 기업의 이미지, 공인에 대한 긍정.부정의 글을 개진할 수 있게 되었다. 이러한 현상에 따라 기업, 제품, 혹은 공공의 분야에서 일반 개인들이 어떻게 생각하는가에 대한 분석 및 자료수집의 필요성이 높아지고 있다. 감성용어 문서분류시스템은 문서의 내용 중 감성기반의 용어들에 기반하여 이에 대한 패턴을 정의하고 이에 대응하는 범주에 문서를 자동으로 할당하는 작업으로써 효율적인 정보 관리 및 검색을 가능하게 한다. 하지만 자동문서 분류를 하기 위해서는 방대한 양의 데이터를 수집 보관하기 위한 분산 환경이 반드시 필요하다. 본 논문에서는 감성기반 문서분류 시스템을 위한 감성용어 추출 및 긍정,부정의 패턴을 검색해 자동 문서 분류를 위해 RTI(Run Time Infrastructure)를 통한 분산 시스템 환경으로 구성하였다.

1. Introduction

최근에는 신문이나 잡지와 같은 미디어부터 인터넷을 이용한 전자매체까지 다양한 경로를 통해 정보를 습득할 수 있게 되었다.

특히, 인터넷의 확산과 더불어 전자매체를 이용함으로써 방대한 양의 정보를 통합하여 사용자에게 제공함으로써 보다 편리하게 정보를 얻고 활용할 수 있게 되었다. 이와 더불어 사용자들은 얻어진 정보를 통해 사용자 본인들의 의견을 개진하는 쌍방향 의사 소통이 확대되고 있다. 온라인 쇼핑몰의 제품평, 소비자 보호센터의 불만글, 각 기업 사이트 혹은 포탈에 올리는 댓글들이 사용자의 의견은 기업의 이미지 관리 및 제품 개진에 대한 창구 역할을 하는 것이다. 이 같은 다양한 창구를 통해 기업, 혹은 제품, 정치적 사안 등에 대한 긍정, 부정 분위기를 파악 할 수 있게 되었다.

감성기반의 자동 문서 범주화는 대량의 문서에서 글쓴이의 감정, 사안 혹은 제품 기업이나 인물의 이미지에 대한 긍정, 부정의 감성을 효율적으로 관리하고 검색하는 것은 방대한 양의 수작업을 필요로 하여 막대한 시간이 소요된다. 이러한 자연언어 처리 시스템에서 기존의 웹 문서 범주를 결정해 주는 문서 분류 시스템에 대한 많은 연구가 있으나 [1,2,3] 감성을 기반으로 자동으로 범주화 및 분류를 시행하는 시스템에 대한 연구는 미흡하다.

본 논문에서는 다양한 문서에서의 감성기반 어휘 및 패턴을 추출 이를 비교하여 문서의 범주 (긍정, 부정)의 여부를 자동으로 분류하는 시스템을 설계, 구축하여 분류/관리 효율성을 도모 하고자 한다.

2. Related Works

2.1 문서 분류의 개관

자동문서분류는 전처리 과정과 감성 용어 축소 과정을 거쳐 문서에 출현하는 어휘들을 바탕으로 문서를 재형성한다. 전처리 과정은 문서로부터 태그와 본문과 관계없는 불용어를 제거하고 형태소 분석을 통해 특정 용어들을 추출하게 된다. 주제어축소과정은 형태소분석을 통해 나온 특징들 중 감성용어에 해당하는 용어들을 추출하는 과정과 함께 가중치를 결정하기 쉽게 하기 위해 특정방법으로 정규화하는 과정을 말한다.

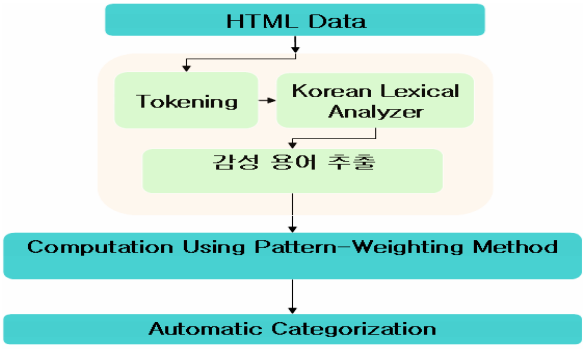

그림 1. 자동문서분류 모델

[그림 1]은 전체적인 감성 기반 자동 문서분류를 모델화한 것이다.

2.1.1 자동 문서 분류 방법

기존의 문서분류 방식은 크게 규칙, 확률벡터, 관련도 방식으로 나뉜다[2]. 규칙기반은 문서 정보의 패턴을 규칙으로 하여 문서를 분류하는 방법을 말한다. 확률벡터 방식[1]은 문서에 나타난 용어들의 출현 분포를 이용하여 문서를 분류하는 방법으로써 분석된 문서로부터 또 다른 용어 분포의 확률을 습득하여 시스템의 분류 규칙을 확장해 나갈 수 가 있다. 하지만 이 방식은 용어간 상관관계를 나타내기는 힘들다. 관련도 기반방식[3]은 분류된 문사와 새로운 문서 사이의 관련도를 찾아 분류하는 방식을 말한다. 하지만 이 방식은 관련도 측정에 있어서 오버헤드가 크다는 단점 역시 가지고 있다. 본 논문에서 감성 어휘와 문서상에서 나타내는 패턴을 활용한 시스템을 작성하기 위해서는 규칙방식을 채택한다. 하지만 시스템의 확장성을 고려하여 패턴위주가 아닌 감성 어휘들의 상관관계에 대한 학습효과가 있고 분석 오버헤드가

적은 신경망 방식을 접목시킨 감성기반 자동 분서 분류 시스템을 설계 구축한다. 또한 감성 용어에 대한 사전을 구축하고, 웹 문서의 가중치를 계산하는 방법으로 이 가중치 계산방법을 적용과 함께 감성문서의 패턴을 적용하여 문서의 감성정보 분석을 수행하였다.

신경망 방식은 training 문서와 범주벡터로 학습 후 새로운 문서에 대해 범주 작업을 시행하며 이에 대한 오류 역시 다시 망의 학습체계로 입력되어 학습패턴 및 범주에 대한 정확도를 증가시키는 방식을 말한다[18][19].

2.2 RTI (Run Time Infrastructure)

RTI는 운영체제와 federate들 사이에 있는 미들웨어 소프트웨어로 각 federate들의 데이터 교환 및 시뮬레이션 시간 진행에 필요한 여러 가지 기능을 제공한다. RTI는 구성 요소는 다음과 같다. [7].

(1) RTIExec (RTI Executive Process)

Federation 실행의 생성과 소멸을 관리하는 전역 프로세스로, FedExec가 서로 다른 이름을 갖도록 하며 수동 조작을 위한 인터페이스를 제공하는 역할을 수행한다.

(2) FedExec (Federation Executive process)

실행중인 federation에서 생성되는 하나의 프로세스로, 생성된 federation을 관리하면 federation에 참가하는 federate들에게 핸들을 할당하여 federation 실행에 참가하고 탈퇴하는 것을 관리한다.

(3)libRTI (RTI Library)

C++ 라이브러리인 libRTI는 응용 프로그램에 포함되는 클래스 라이브러리로, 각각의 응용 프로그램들을 libRTI를 통해서 RTI의 서비스를 호출할 수 있다[8].

Federation의 실행은 그림 2처럼 3단계로 수행된다. 각 단계는 아래와 같다.

(1) 사용자는 RTI를 시작하여 RTIExec 프로세스를 실행시켜 새로운 federate가 참가 가능 하도 록 준비한다.
(2) 사용자는 federate를 실행하고 이 federate는 새로운 federation을 만들며 FedExec 프로세스를 실행시킨다.
(3) 새롭게 추가 되어지는 federate들은 FedExec를 통하여 기존의 federation에 참여하게 된다.

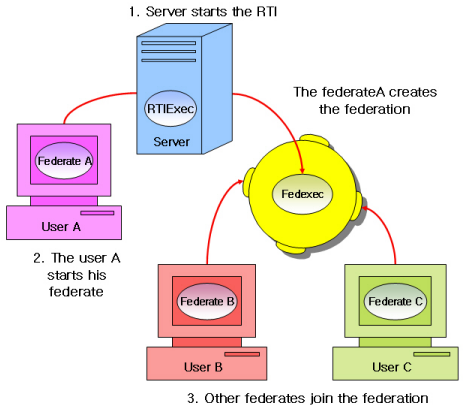

그림 2. RTI federation의 진행

RTI, 상호 운용성을 위한 본질적인 방법을 제공한다. 이것은 federate들과 federation간의 정보교환을 가능하게 하고 실시간 데이터 전송에 관한 수단과 개념을 제공한다[12] [13] [14].

3. 실험환경 모델

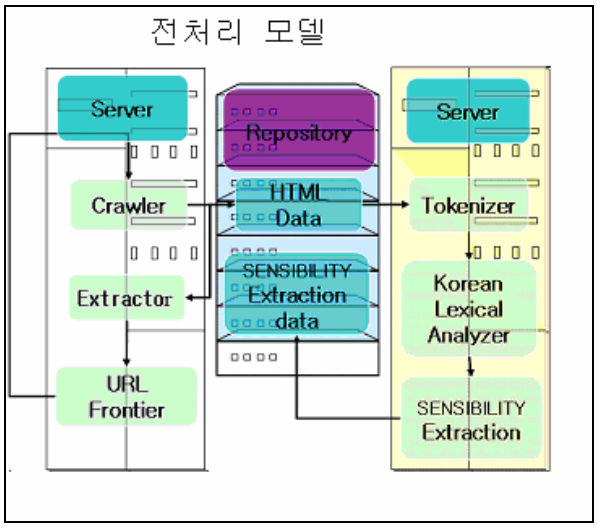

그림 3. Susceptibility Extraction System

그림3은 자동 문서분류 과정 중 첫 번째 단계로 주제어추출모델을 나타낸다. 한글 웹 문서를 대상으로 각 문서의 감성용어를 추출해낸다.

3.1.1 Crawler

본 논문에 사용된 크로울러는 **seed URL** 주소로부터 시작하여 자원이 사용 가능한 한 계속적으로 페이지들을 방문하고 새로운 주소를 추출 후 넘겨받아 넘겨받은 주소를 다시 방문하는 과정을 반복한다[15]

Extractor는 다운 받은 html 파일로부터 Child URL들을 추출하게 되고 Frontier에 중복된 URL이 있는 지 확인하는 DUE를 통해, 없으면 URL Frontier에 Child URL들을 저장하게 된다.[9,10,11]

3.1.2 Korean Lexical Analyzer

Token화된 웹 문서에 대해서 그 문서를 대표 할 수 있는 한글 주제어들을 추출하기 위해 형태소분석기(HAM)를 사용하였다.

한국어는 영어와 달리 한 어절이 하나 이상의 형태소로 이루어져 있기 때문에 주제어를 추출 하기 위해서는 어절을 이루고 있는 형태소들에 대한 인식 및 분리가 이루어져야 한다.[6]

3.1.3 Susceptibility Extraction, Susceptibility Extraction data

Susceptibility Extraction은 형태소 분석한 결과 에서 감성용어들을 추출해내는 역할을 한다. 이후 감성관련용어만으로 이루어진 파일을 Susceptibility Extraction data에 저장하게 된다. 용어 추출 시에는 감성용어 사전을 통한 비교분석을 통해 추출한다.

용어 추출방식은 먼저 수집한 문서의 대 범주, 정치, 경제, IT등의 문서를 분류하기 위한 전문용어 사전을 통한 문서분류 후 다음 단에서 설명하는 Categorizator에 감성용어 및 패턴의 매칭을 위한 입력으로 넘겨진다.

3.2 Categorizator

Categorizator는 가공된 각 한글 웹문서와 구축된 감성용어 사전 및 패턴을 통해서 각 문서의 Frequency를 계산하여, 감성에 해당하는 문서로 분류하는 기능을 한다. 또한 범주의 계산과 학습을 위해 오류 후진전파[16][17]알고리즘을 통한 방식을 사용하였다.

그림4는 앞 절에서 설명한 입력을 자동 분류하는 시스템의 처리 관계도이다.

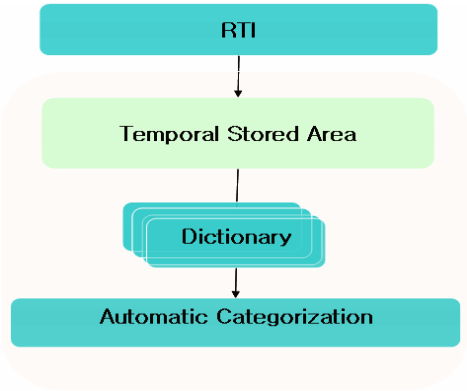

그림 4. **Categorizator Architecture**

3.3 RTI 기반 분산환경 구축

앞서 얘기한 전처리 시스템부터 감성용어 추출까지의 과정을 거친 문서들은 단일 시스템이 아닌 분산 시스템에 그 데이터를 나누어서 관리하게 된다. 이 과정에서 각 분산 시스템에서 감성관련 사전 및 패턴을 가지고 있는 agent를 통해서 들어오는 데이터의 범주를 계산하여 해당 주제의 문서정보만을 모아서 관리하게 된다. 이때 각 federate들은 다운 받은 웹 페이지를 관리 하는 것이 아닌 웹 페이지가 저장장치인 RAID 어느 곳에 위치해있는지에 관한 위치정보와 페이지의 범주를 계산하기 위한 추출된 주제어 정보만을 받게 된다.

각 문서의 분류 시 통신을 위해 RTI를 활용 한다. RTI는 앞서도 말했듯이 federation간의 실시간 데이터 전송에 관한 수단 과 개념을 제공한다 [12] [13] [14]. 그림 5는 이상의 시스템의 전체적인 구성을 보여준다.

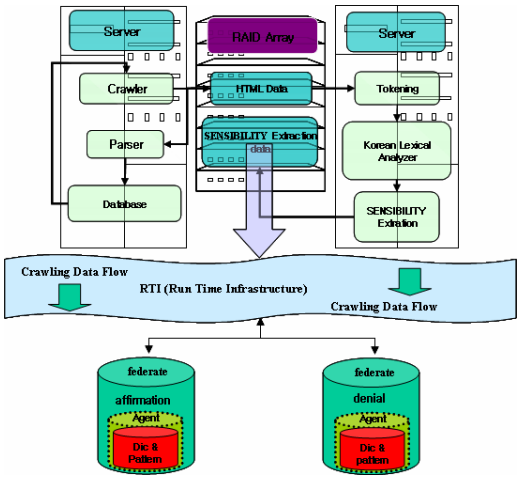

그림 5. Distributed Document Management System with RTI

4. 구현 및 실험 분석

4.1 실험 방법 및 데이터 구성

실험은 감성용어 사전 및 패턴을 통한 문서 분류를 함에 있어서 긍정, 부정의 사전 및 패턴을 agent화 하여 각 agent에서 자동으로 해당 감성을 걸러내는 것을 실험 하였다. 문서 분류의 실험 문서집단으로서는 “소비자가 만드는 신문”, “LG, 삼성의 고객센터”에서 총 710건의 문서를 사용하였다. 이중 긍정 150 부정 150건으로 사전 구성 및 패턴 정의를 하였으며, 사전학습세트를 제외한 긍정 160건, 부정 250건을 테스트 세트로 하였다.

표1은 감성용어의 패턴이 문서에서 나타나는 양상의 예를 보인 것이다. 실제 패턴의 검색을 위해서 표와 같은 양식으로 매칭 검사를 하였다.

표 1. 감성 용어 패턴정의 예

	선행	후행
긍정	불구하고 근데 하지만 불편 • • •	친절 흔쾌히 감명, 정확한, 신속히, 뿌듯, 잘해주셔서(잘해주다) • • •
부정	했습니다. 물었습니다 소비자 저뿐만 뭔가 이상하다 아직 • • •	그런데 소비자과실 손해 당하는 역시 횡포 • • •

4.2 RTI통한 분산처리

다음의 그림은 RTI를 통한 federation들의 연결 화면이다.

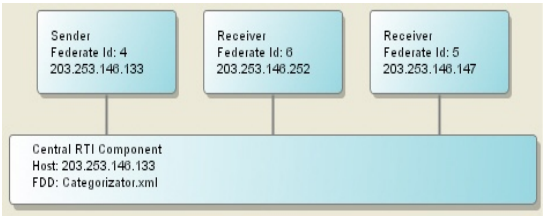

그림 6. Federation Connection

각 federation들은 Crawler서버인 Sender로부터 웹데이터의 위치와 주제어 정보를 받는다. 이후 각 federation에 설치된 Categorizator를 통해 해당 주제의 분류를 시행 federation별 해당 주제에 관한 데이터만을 저장 관리 한다. federation에서 다루지 않는 주제에 대해서는 federation에서는 garbarge처리를 통해 관리 대상에서 제외 시킨다.

[그림 7]은 긍정에 관한 사전 및 패턴을 적용했을 때의 문서 분류 결과이고 [그림 7.a]는 부정사전 및 패턴을 적용해 분류한 결과 화면 이다.

표 2는 각 agent 에서 긍정, 부정의 결과를 분류한 결과를 나타낸 것이다. 무시된 항목은 사전 적용 결과 agent에서 관리하는 주제 분류 에서 해당 조건에 만족하지 않는 것으로 이는 Garabage 항목에도 해당하지 않는다.

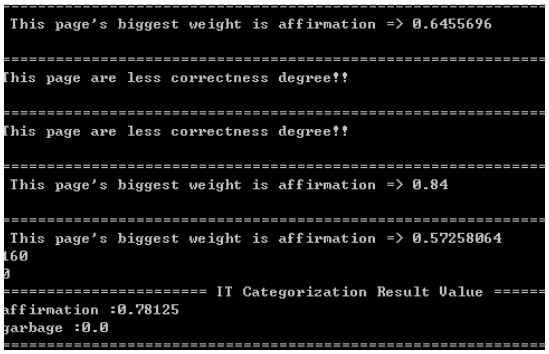

그림 7. 긍정 범주화 계산 결과

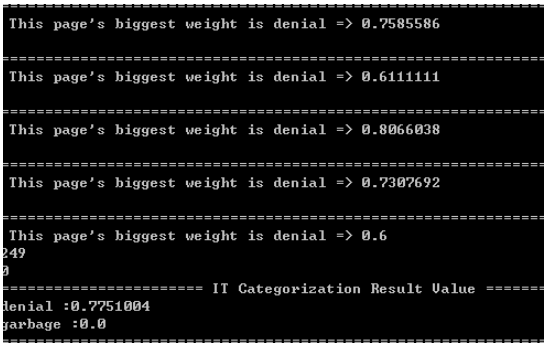

그림 7.a. 부정 범주화 계산 결과

표 2 정규사전과 기사를 통해 구축한 사전의 분류 결과

	긍정	부정
Federation 1	125	0
Federation 2	0	193

긍정주제를 가진 agent 에서는 총 410 건의 데이터를 받아들인 후 이중 125 건을 긍정으로 분류하였다. 긍정이 아니라고 판단한 문서에는 부정글 250 건과 긍정글 35 건이 포함되었다. 이는 긍정이란 주제로 분류하기 위해 사전 및 패턴매칭 결과 0.5 의 가중치에 해당되지 않은 문서들이다. 마찬가지로 부정주제의 agent 에서는 총 410 건의 문서 중 긍정글 160 건과 부정글 57 건이 포함되었는데 제외 이유는 긍정 agent 에서와 같은 이유다.

그림 8은 긍정, 부정 이외에 중립적인 입장과 같은 또 다른 감성에 대한 구분을 위해 RTI를 통한 agent확장화면이다. 임의의 장비에 분류하고자 하는 주제의 사전을 구축하고 RTI federation에 연결만 함으로써 추가적인 주제의 문서 분류를 시행 함을 확인 할 수 있다.

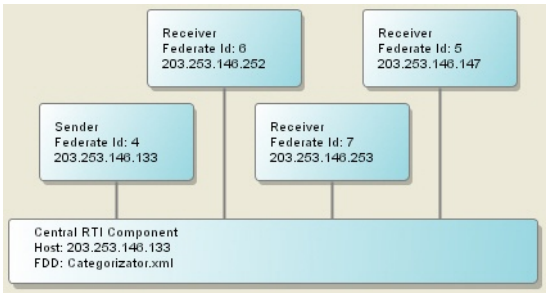

그림 8. Extend Federation Connection

5. 결론 및 Future Work

본 논문에서는 감성용어 사전을 구축하는데 기반이 되는 웹 글에 대한 분석을 하고 그 결과를 토대로 사전 및 패턴을 만들어 각 federation에 agent 설치 이를 바탕으로 주제별 분산 분류를 시행 했다. 각 federation들은 RTI를 통해 웹 페이지의 저장장치에서의 위치정보, 그리고 해당 문서의 주제어를 축소한 결과 만을 받아서 이를 통해 갖추어진 사전과 비교해 문서분류를 분산 처리를 할 수 있었다. 각 federation들에게는 파일 전체가 아닌 비교에 필요한 요약 정보만을 줌으로써 각 federation이 파일을 가지고 직접 비교 할 때의 경우보다 많은 로드를 줄일 수 있다. 또한 모든 실제 데이터는 RAID array한곳에서 관리하기에 각 federation들의 디스크공간의 중복 낭비 방지의 효과도 기대할 수 있다.

전제적인 시스템에서 문서분류의 정확성은 각 federation들의 사전 및 패턴을 얼마나 정확성을 높여서 구축하느냐에 있다.

향후 구축에 필요한 범주간의 감성 용어정보에 대한 분석을 좀더 정확성 있게 시행하고, 감성 패턴에 대한 정확한 규정을 구축한다면 보다 효율적인 자동분류를 수행할 수 있는 시스템을 만들 수 있을 것으로 판단된다.

<Acknowledgement>

본 논문은 산업자원부 한국산업기술평가원 지정 한국항공대학교 부설 인터넷정보 검색 연구센터의 지원에 의함

References

[1] K.M Wong and Ziarko, V.V.Raghavan, and P.C.N. Wong, "On Extending the Vector Space Model for Boolen Query Processing," In Proc. Intl. Conf. On Research and Development in Information Retrieval, ACM SIGIR, pages 175-185,1986

[2] 권오욱, 확률벡터와 메타범주를 이용한 최적 문서 범주화 모델, 석사학위논문, 한국 과학 기술원 전산학과, 1995

[3] Yang, "Expert NetWork: Effective and Efficient Learning from Human Decision in Text Categorization and Retrieval", In Proc. Intl. Conf. on Research and Development in Information Retrieval. ACM SIGIR, pages 13-22, 1944.

[4] Bao, Y. and Ishii, N., "Combining Multiple K-Nearest Neighbor Classifiers for Text Classification by Reducts", In Proceeding of the fifth International Conference on Discovery Science, 2002

[5] Sasaki, M. and Kita, K., "Rule-Based Text Categorization Using Hierarchical Categories", In Proceeding of the IEEE International Conference on System, Man and Cybernetics, 1998

[6] Sebastiani, F., "Machine learning in automated text categorization", ACM Computing Survwys, 2002

[7] 서혜숙, 한상범, 신종회, 황종선, “XML을 적용한 HLA 기반 시뮬레이션” 한국 정보 과학회 2003 춘계학술발표논문집, pp.115-117
[8] 차영필, 정무영, “분산 생산 시스템을 위한 agent기반의 협업 시뮬레이션 체계” 2003 춘계공동학술대회.
[9] Jaeyun, lee., Boyeong, Choi., yeongmi, Jeong., " Research about term weight’s techniques in literature automatic categorization", Korea Society for Information Management, 2000
[10] Gwangje, Jo., Juntae, Kim., "Research document’s automatic categorization in hierarchic category system by Inverted Term Frequency", Korea Information Science Society, 1997
[11] Marc Najork and Allan Heydon, “High-Performance Web Crawling”, 2001
[12] IEEE 1516, "Standard for Modeling and Simulation High Level Architecture - Framework and Rules" IEEE, 2000.
[13] IEEE 1516.2, "Standard for Modeling and Simulation High Level Architecture - Federate Interface Specification" IEEE, 2000.
[14] IEEE 1516.1, "Standard for Modeling and Simulation High Level Architecture - Object Model Template Specification" IEEE, 2000.
[15] A. Arasu, J.Cho, H.Garcia-Molina, A.Paepcke, S.Raghavan, "Searching the Web", ACM Transactions on Iternet Technology, Vol,1, Num. 1, August 2001, pp.2-43.
[16] D.E.Rumelhart, G.E.Hinton, and R.J.Williams, “Learning Internal Representations by Error Propagation”, Parallel Distributed Processing, Vol.1, 1986
[17] Philip N. Johnson-Laird “The Computer and the Mind: An Introduction to Cognitive Science”, Harvard Univ. Press, 1988
[18] http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html
[19] http://www.cs.toronto.edu/~hinton/csc321/lectu re s.html