WWW상의 지능형 정보검색을 위한 기계학습 알고리즘 구현에 관한 연구*

A Study on Machine Learning Algorithm for Intelligent Information Retrieval in World Wide Web

김성희(Seong-Hee Kim)**

목 차

초 록

본 연구에서는 현재 웹정보검색의 문제점을 해결하기 위하여 기계학습알고리즘을 이용한 지능형정보검색 시스템을 구현하고 있다. 구체적으로, 수학분야 질의어 및 적합한 문서를 선정해서 이 자료를 토대로 어떻게 귀납학습알고리즘과 신경망을 적용할 수 있는지를 검토하고 있다. 또한, 본 논문에서는 신경망시스템 설계시 성능에 영향을 미치는 입.출력노드수, 은닉층, 학습매개변수등 다양한 요소를 경험적방법을 통해 검토하고 있다.

ABSTRACT

We investigate the appropriate design and implementation of an Inductive Learning Alogrithm with a Neural Network in order to solve both inconsistent indexing and incomplete query problems on the web. Specifically, the proposed system based queries and documents in the field of Mathematics shows how inductive learning method and neural networks can apply to information retreival. Also, this study examines all of parameters of the neural networks -- the number of node in input and output, hidden layer size and learning parameters etc. -- which are significant in determining how well the neural network will converge.

키워드: 인터넷, 웹정보검색, 지능형정보검색, 기계 학습 알고리즘, 검색효율성

* 이 논문은 1998년 한국학술진흥재단의 학술연구비에 의하여 지원되었음
** 동덕여자대학교 정보대학 정보학부 조교수
■ 논문 접수일 : 2000년 5월 18일

1 서 론

최근 들어 컴퓨터 성능의 향상과 통신기술의 발달로 인터넷 이용자가 급속히 증가하고 있다. 최근 들어서는, 인터넷에 존재하는 일반 text 형태의 문서, 그림, 오디오등의 각종 데이터를 URL(Uniform Resource Locater: 인터넷에 있는 임의의 정보의 주소를 지정하는 방식) 를 이용하여 하나의 문서 형태로 통합적으로 제공한다. 대표적인 서비스로는 WWW(World Wide Web)로 이는 hypertext에 기반한 인터넷 서비스이다. WWW의 사용이유는 hypertext 기사 읽기와 인터넷 자원에 접속하는 것이라고 할 수 있다. 첫째, hypertext가 보통 일반 문서와 다른 점은 문서들이 연결되어 있다는 것이다. 문제는 hypertext 안에 연결관계를 설정하는 것은 시간을 많이 소모한다는 점과 hypertext 연결이 얼마나 적절한가 하는 점이다. Hypertext 문서는 hypertext 이용자의 생각과 그 문서에 연결시키는 사람과 얼마나 밀접한가에 따라 hypertext가 유용할 수 도 있고 그렇지 않을 수도 있다는 것이다. 둘째, WWW를 이용하는 또 다른 이유는 WWW상에서 다른 다양한 인터넷 서비스에 접근할 수 있다는 것이다. 예를 들면, WWW를 통해, telnet, gopher, ftp, E-mail과 같은 서비스에 접속을 할 수 있다. 이와 같이 WWW는 인터넷에서 제공하고 있는 여러 가지 서비스를 하나로 통합하여 사용할 수 있는 장점을 제공하므로 인터넷 이용은 기하 급수적으로 증가하고 있다.

한편, 인터넷사용이 급증과 더불어, 인터넷 정보 또한 급증하면서 이용자들이 원하는 정보를 신속하고, 정확하게 제공할 수 있도록 도와주는 탐색엔진들이 제공되어 왔다.대표적인 WWW 탐색엔진으로는 야후(Yahoo), 알타비스타(Altavista), 라이코스(Lycos), 익사이트(excite), 인포시크(InfoSeek), 심마니, 까치네 등을 들 수 있다. 그러나 이런 많은 탐색엔진들은 각자, 데이터베이스 구축방법, 탐색기법, 검색대상, 출력 내용 등이 다르기 때문에, 검색결과가 다양한 것으로나타났다(Courtois et al., 1995). 또한 이들 탐색엔진들은 서로 다른 검색결과의 제공 이외에도 상이한 사용자 인터페이스로 인하여 사용자가 원하는 모든 정보 검색은 어렵게 되었다. 이런 이유로 사용자가 쉽게 사용할 수 있는 단일화된 인터페이스를 제공하며 여러 탐색엔진의 검색 결과를 종합해줄 수 있는 메타탐색엔진이 등장하게 되었다. 대표적인 메타 탐색엔진으로는 MetaCrawler, Savvy Search, All in One, Internet Sleuth 등이 있으며, 국내에서는 미스 다찾니가 있다.

이상에서와 같이 인터넷 상에 정보의 양과 종류, 정보제공자의 수가 급격히 증가하면서 인터넷 사용자들은 정보검색도구로서 탐색엔진(search engine)을 많이 이용하고 있다. 특히, 최근 들어 hypertext형태의 정보조직 및 브라우징(browsing)이 가능한 WWW(World Wide Web)를 이용하여 정보를 제공하는 사이트가 기하급수적으로 늘어나면서 WWW를 이용해서 인터넷 상에 널리 퍼져있는 정보를 사용자에게 찾을 수 있도록 도와주는 도구인 탐색엔진들이 개발되고 있는 추세이다. 현재 전 세계적으로 많은 탐색엔진들이 이용되고 있고, 각종 정보에 대한 데이터베이스를 구축하고 계속적으로 갱신함으로써 사용자에게 최신의 정보를 제공하고 있다. 탐색엔진은 로봇 에이전트(robot agents)라는 웹을 순회(traverse)하는 프로그

램에 의해 정보를 수집한다. 로봇 에이전트는 지정한 URL(Uniform Resource Locator)을 순회하며 각 홈페이지들의 정보를 수집하고 수집된 정보에 대해서 탐색엔진은 사용자가 편리한 방법으로 검색할 수 있는 인덱스를 제공한다. 현재 대표적인 WWW탐색엔진은 WWW의 대부분의 정보가 HTML문서이고 HTML문서가 텍스트라는 점에서 기존의 정보검색(Information retrieval)기술을 WWW에 적용한 시스템이라고 할 수 있다. 따라서 WWW탐색엔진은 WWW상의 HTML문서를 색인하여 이 색인을 이용자의 질의어와 비교하여 적합한 HTML문서들의 URL을 제공해준다.

그러나 하루가 다르게 커지는 WWW, 더욱 다양해지고 있는 WWW상의 정보의 내용, 정보의 유형, 그리고 동적인 HTML문서를 지원하는 다양한 웹 애플리케이션(Web application)의 증가로 기존의 WWW검색도구들인 웹탐색엔진에는 검색도구로서의 장점에도 불구하고 다음과 같은 단점이 있다.

- WWW상의 HTML문서는 자주 내용이 바뀌고 문서자체가 삭제되고 새로운 HTML문서가 추가된다.
- 모든 웹문서를 색인 하는 것은 거의 불가능하다. 예를 들면, 다른 HTML문서에서 참조되지 않는 잘 알려지지 않는 웹 서버를 색인 하는 것은 거의 불가능하다.
- 적절한 질의어를 작성하는데 어려움이 있다. 이는 전통적인 정보검색의 문제로 잘 알려져 있다.

이상과 같이 WWW탐색엔진은 색인의 불완전성, 부정확성, 그리고 이용자 질의의 부정확성등으로 인해 질의어와 색인어의 불일치문제가 있어 원하는 정보를 검색하는데 실패하는 경우가 많다. 이런 경우 검색하고자하는 정보가 실제로 웹상에 존재하지 않는 것인지 아니면 존재하는데 찾지 못하는 것인지 이용자는 판단할 수 없다. 따라서 지금의 탐색엔진만으로는 원하는 정보를 검색하는데 문제가 있다. 이런 문제점을 해결하기 위해 최근 들어 새로운 검색들이 연구되었다. 예를 들면, 인터넷상의 정보를 사용자의 관심에 따라 필터링해서 이용자의 관심에 부합하는 정보만을 사용자에게 제공해주는 filtering system, 이용자가 관심 있는 웹사이트를 정기적으로 방문, 확인하여 변경된 사항을 알려주는 monitoring system등이 있다. 그러나 이런 연구결과들은 현재 웹탐색엔진의 문제점을 부분적으로 해결할 수 있지만 근본적인 해결책은 되지 못하고 있다. 즉, 이들 모델은 현존하는 불리언 정보검색의 단점을 보완할 수 있다는 것을 증명해왔으나 이들 모델 역시 상징적이며 본문일치수준(text-match level)에서 머무는 키워드 검색에 기초하고 있으며, 검색과정에서의 의미론적이고 문맥적인 정보를 무시하고 있다(Watter 1989). 따라서, 기하급수적으로 증가하는 인터넷상의 흩어져있는 정보를 이용자가 원하는 정보만을 골라 검색할 수 있는 새로운 연구가 절실히 필요하다. 따라서 본 연구에서는 먼저, 현재 웹상의 다양한 형태의 정보검색의 문제점을 해결하기 위해 선행연구를 조사한 후 기계학습알고리즘을 적용한 지능형정보검색 시스템을 하나의 prototype으로 구현하고자 한다. 즉, 신경망(neural network) 및 귀납학습(inductive learning)을 이용함으로써 이용자가 찾고자 하는 정보를 정확하게 표현하지 못하거나 또는 문헌을 표현하는 색인이 일관성이 없을 경우에도 이용자가 원하는 정보를 검색하여 제공함으

로써 이상적인 지능형 정보검색시스템을 개발하는 기초를 제공하고자 한다.

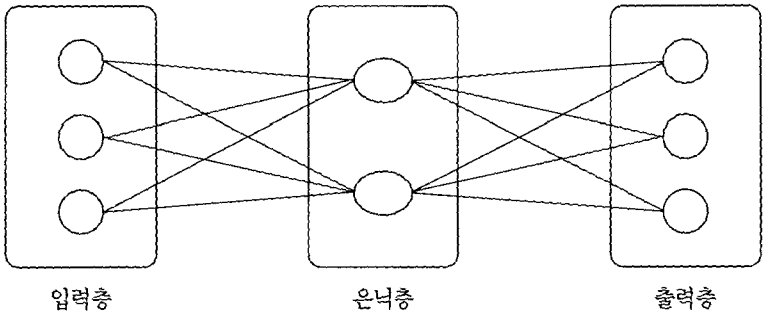
2 기계학습

기계학습은 하나의 학습시스템으로써 기존의 경험들을 학습시킴으로써 새로운 문제가 주어졌을 때 추론을 통해 해결하는 시스템이라 할 수 있다. 대표적인 기계학습 알고리즘으로는 신경망과 귀납학습을 들 수 있다.

신경망(neural networks)은 인간의 뇌 그리고 신경세포가 반응하는 것과 유사하게 설계된 회로이다. 이는 많은 수의 소자를 네트워크로 연결하고, 각 소자들 사이의 연결의 세기(connection strength)로 정보를 표현하고 기억한다. 이는 인간두뇌와 같이 비결정적인 특성을 가지고 있으므로 약간 틀리거나 비슷한 입력을 인식할 수 있다. 음성인식, 문자인식, 영상처리, 자연어 이해 등의 분야에 주로 이용되고 있다. 가장 대표적인 신경망으로는 역전파(back-propagation) 신경망이다. 이것은 입력층(input layer), 출력층(output layer) 그리고 이들 사이에 1개 이상의 은닉층(hidden layer)으로 구성되어 있다 〈그림 1〉.

일반적으로 입력층은 주로 질의어로 구성되고 출력층은 각 질의어에 해당되는 적합한 문서 색인어가 해당이 된다. 또한, 이들간의 은닉층을 통하여 이상적인 시스템을 구현하게 된다.

역전파신경망은 그 동안 정보검색시스템엣 광범위하게 적용되어왔다. 예를 들면 Mital and Gedeon(1991), Gersho and Reiter (1990), 그리고 Wilkinson and Hingston (1992)의 연구들은 모두 신경망이 지능형 정보검색에 응용할 수 있다는 것을 보여주었다. 먼저, Mital and Gedeon(1991)은 정보검색을 위해 은닉층 없이 입력층과 출력층으로만 구성된 신경망을 적용하였다. 여기서 569노드수를 이용하였으며 19개문헌을 사용하였다. 또한, Wilkinson and Hingston(1992)은 9,000개의 노드를 사용하고 입력, 은닉층, 그리고 출력층 등 3개의 층을 사용하였다. 그러나 각층의 구체적인 노드수에 대해서는 언급을 하지 않았다. Gersho and Reiter(1990)는 역전파 신경망을 이용한 다층 신경망 네트크를 정보검색에 응용하였는데 이들 연구에서는 특히, 알고리즘성능



〈그림 1〉 신경망 예

에 영향을 미치는 요인인 학습매개변수와 모멘텀에 대한 언급을 하였다. 자세한 역전파신경망에 대한 이론적 설명은 김성희(1994, 1999)논문에 설명이 되어있다. 한편 귀납학습(inductive learning)은 training examples로부터 추론을 통해 지식을 획득하는 과정이라고 할 수 있다. 일반적으로 귀납학습알고리즘은 크게 입력과 출력부분으로 구성된다. 입력부분은 다시 3가지 부분으로 구성된다: (1) training examples 세트, (2) 생성 규칙(generalization and rules), 그리고 (3) 추론기준(criteria for inference).여기서 training examples는 두 개의 부분으로 세분할 수 있는데, attribute 세트와 classification decision이다. 예를 들면, 본 논문에서는 각 질의어 색인어가 attribute set이 되고 이들 질의어에 대한 적합한 문서의 색인어가 classification decision이 되는 것이다. 귀납학습알고리즘의 출력부분은 결정규칙(decision rules)이다. 즉, 각 질의어에 대한 적합문서를 검색하기 위한 의사결정트리를 생성하는 것이다.

그 동안 귀납학습에 대한 프로그램이 많이 개발되어왔다. 대표적인 것으로는 AQ-Star(Michalski 1983), PLS(Rendell 1986), ID3 (Quinlan 1986)등이 있다. 본 논문에서 사용된 프로그램은 ID3의 새로운 버전인 C4.5이다.

C4.5는 information gain ratio criterion을 사용하면서 의사결정트리를 생성한다.예를 들면, S개의 training examples가 있다고 하고 이 S개의 examples중에 p positive examples와 n negative examples가 있을 경우 그 엔트로피는 H(S)로 다음과 같이 정의할 수 있다.

$$H(S)=-p/(p+n)\log 2p/p+n)$$

$$-\ n/(p+n)\log 2n/(p+n).$$

본 논문에서는 귀납학습알고리즘을 색인의 불일치성의 문제점을 해결하는데 적용하고 있다. 즉, 각 질의어에 대한 적합문서를 positive examples로하고 부적합한 문서를 negative examples로 구분해서 학습시킴으로써 질의어에 대한 적합한 문서색인어를 추출할 수 있는 의사결정트리구조를 생성하는 것이다. 구체적인 설계는 시스템 구현부분에서 설명하고 있다.

3 선행연구

인터넷사용의 급증과 WWW상의 폭발적인 정보의 증가는 이용자들로 하여금 원하는 정보를 검색하는데 어려움을 주고 있다. WWW탐색엔진이 매우 유용한 도구이지만 웹탐색엔진만으로는 이용자가 원하는 정보검색이 어렵다. 웹 탐색엔진은 기존의 정보검색 시스템의 문제점을 그대로 갖고 있을 뿐 아니라 WWW상의 문서는 통제가 불가능한 분산 시스템이고 정보의 내용이 계속 변하기 때문에 색인을 완벽하게 구축한다는 것이 근본적으로 불가능하다. 따라서, 전통적인 정보검색의 문제점과 마찬가지로 WWW탐색엔진의 문제점은 이용자의 질의어와 HTML문서의 색인어사이의 불일치성이다. 최근 들어 WWW 정보검색의 문제점들을 해결하기 위한 연구가 상당히 활발하게 이루어져 왔다.

먼저 조강래(1997)는 인터넷 이용자가 관심있는 웹사이트 리스트를 미리 지정해서 이 웹사이트를 정기적으로 방문해서 갱신된 내용이 있는지 확인해서 변경된 사항을 이용자에게 알

려주는 모니터링 시스템(monitoring system)을 제안하였다.

오종인 등(1997)은 웹정보검색 시스템의 정확도 향상을 위해 상대 출현빈도와 역문헌 빈도에 의해 추출된 색인어에 대하여 원문을 대상을 공기(co-occurrence)단어들을 추출하고, 상호정보량의 계산에 의해 색인어들을 의미에 따라 재분류함으로써 재현률을 감소시키지 않고 정확도를 향상시킬 수 있는지를 실험하였다. 그 결과 별도의 의미 사전이나 태그 없이도 정확도를 4.5% 정도 향상시킬 수 있다는 결론을 내렸다. 이는 색인어의 의미분류는 웹 정보검색에 중요한 요소임을 입증하였다고 볼 수 있다.

조영재 등(1997)은 웹 에이전트를 이용한 효과적인 정보검색방안을 제안했다. 그들에 따르면, 웹 에이전트는 사용자가 원하는 정보의 성향을 연관 피드백을 통해 학습하고 학습된 내용에 따라 웹 상의 정보들을 필터링 함으로써 정보검색 시스템의 효율성을 증진시킬 수 있다고 하였다.

Goldberg(1992)는 최초로 Tapestry라는 메일추천시스템을 개발하였다. 이 연구에 따르면 비슷한 관심분야를 갖는 사람들이 서로 협동하여 필요한 정보를 기록하였다가 비슷한 검색이 인터넷 상에 발생했을 때 이 기록된 정보를 이용함으로써 문서를 직접 읽지 않고도 적합성 판단을 할 수 있게 한다는 것이다. 이런 시스템은 Usenet News에 적합한 것으로 보인다. 현재 수많은 뉴 스그룹이 있으며 각 뉴스 그룹마다 매일 수많은 뉴스가 포스팅(posting)된다. 검색 대상을 몇 개의 뉴스 그룹으로 제한한다 하더라도 많은 뉴스들을 일일이 읽고서 적합한 정보를 검색하는 일은 어렵기 때문에 유용한 정보들만을 선택해서 추천해주는 시스템이 유용할 것으로 보인다.

Yan(1995)과 Sheth(1993, 1994)는 인터넷 상에서 매일 계속적으로 들어오는 수많은 디지털 정보들을 필터링해서 이용자에게 필요한 정보만을 선택적으로 전달해줄 수 있는 필터링 시스템을 제안하였다. 그러나 이런 필터링 시스템의 경우, 정보의 유입창구가 분산되어 있고 어디로 유입되는지에 대한 정보도 없으며 일일이 확인하기에는 인터넷의 크기가 너무 광범위하기 때문에 인터넷에서의 필터링 시스템은 주로 유입창구가 정해져 있는 UsenetNews, E-mail등을 정보원으로 제한하는 경우가 대부분이었다.

이상에서 기존의 WWW 탐색엔진의 문제점들을 해결하기 위해 연구한 결과들에 대해 살펴보았는데 이들 연구들은 근본적인 WWW탐색엔진의 문제점을 해결하지는 못하였고 단지 기존의 WWW탐색엔진을 개선, 보조하는 수단으로 제공되어질 수 있다.따라서 자주 변하는 WWW상의 정보, 그리고 한곳에 집중되어 있는 것이 아니라 여러 곳에 분산되어 있는 정보들을 이용자들이 원하는 정보만을 검색할 수 있는 새로운 검색시스템이 개발되어 져야 할 것이다.

4 지능형 정보검색시스템 설계 및 구현

본 연구는 최근 들어 기하급수적으로 증가하고 있는 WWW상의 정보 중에 이용자가 원하는 정보만을 빠르고 정확하게 제공할 수 있도록 기계학습알고리즘을 이용한 지능형 검색시스템을 설계, 구축, 테스트하기 위한 것이다. 구체적인 구현절차는 다음과 같다.

제 1단계: 문헌타이틀로부터 색인어를 추출한다.

제 2단계: 귀납학습알고리즘을 이용해서 핵심색인어를 추출한다.

제 3단계: 질의어 세트로부터 질의어 색인어를 추출한다.

제 4단계: 질의어와 문헌 핵심색인어를 신경망을 이용해서 매핑시킨다.

제 5단계: 실험질의어를 입력해서 검색효율성을 검토한다.

문헌 및 질의어 선정은 수학관련분야10개의 질의 및 각 질의에 적합한 문서 35개의 문헌 타이틀로 구성되고 있다.이들 질의어 및 문헌타이틀에서 불용어를 제외한 키워드 모두를 색인어로 추출한 후 이러한 질의어 및 문헌 색인어를 학습시켜서 핵심 색인어를 추출한다. 이 때 핵심색인어 추출을 하기 위해귀납학습알고리즘을 이용하고 있다. 사실상 다양한 귀납학습 알고리즘이 있지만 본 연구에서는 Quilan에 의해 최근에 개발된 C4.5방법이 사용될 것이며 이를 수행하기 위해 Sun workstation을 사용할 것이다.또한 질의어와 각 질의어에 대한 적합문서를 mapping시키기 위해 역전파 신경망을 이용할 것이다. 구체적인 설계 및 구현절차는 다음과 같다.

4.1 실험 질의어 및 문헌으로부터 색인어 선정

기계학습은 특정질문에 대한 적합한 문헌을 검색할 수 있도록 학습시키는 것이므로 이런 학습을 시킬 다양한 질의어와 그에 따른 적합한 문서를 미리 설정해야 한다. 본 논문에서는 수학관련분야에서 10개의 질의어 및 35개의 문헌을 사용하고 있다[부록 1]. 이런 질의는 현재 수학분야 박사과정에 재학중인 학생으로부터 전공분야 질의어 및 그에 따른 적합한 문헌을 조사한 것이다. 질의어 및 문헌으로부터 색인어 추출은 불용어를 제외한 모든 단어를 추출하였다. 〈표 1〉과 〈표 2〉는 질의어 및 문서 색인어들이다.

〈표 1〉 문서색인어

advanced,	algebra,	analysis,	analytic,	angular,
application,	approximations,	average,	basis	bergman,
centers,	certain,	characterization	combinatorics,	compact,
complex,	composition,	concepts,	conformal,	contributions,
convex,	convolution,	course,	curves,	derivatives,
differential,	elementary,	essential,	euclidean,	first,
fourier	function,	functional,	function,	geometric,
graph,	gridded,	hardy,	hypergraphs,	inequalities,
integration,	interpolation,	introduction,	lie,	mapping,
mathematical,	maximal,	measure,	methods,	modeling,
molecular,	multiplicate,	norm,	numbers,	numerical,
one,	operators,	over,	plane,	probability,
process,	pseudo,	radial,	rational,	real,
representation,	scattered,	series,	several,	singular,
space,	stochastic,	strongly,	theory,	variable.

〈표 2〉 질의어

질의어 1:	composition,space,operator
질의어 2:	graph, theory,combinatorial,algorithm
질의어 3:	complex,analysis
질의어 4:	series, analysis,fourier
질의어 5:	real, analysis, elementary, contributions
질의어 6:	probability,process,mathematical,
질의어 7:	functional, analysis, multivariate
질의어 8:	numerical,analysis,methods,modeling
질의어 9:	convex, theory, differential,equation,analysis
질의어 10:	representation, theory,methods,rational

4.2 문서 색인어 선정을 위한 귀납학습 알고리즘 적용

현재 WWW 정보검색 문제점중의 하나는 앞에서도 기술하였듯이 웹상의 문서는 자주 수정. 변경. 삭제. 추가되기 때문에 모든 웹상의 문서를 망라해서 색인하기가 어려울 뿐 아니라 색인의 일관성의 문제가 있다. 따라서, 이런 문제는 검색의 효율성을 저하시키므로 이 일관성 없는 색인문제점을 해결하기 위해 귀납학습(inductive learning)알고리즘을 적용한다. 귀납학습은 어떤 사물들을 특정기준에 따라 분류하기 위해 사용되는 인공지능기법중에 하나로써 현재 신경망(neural networks)과 더불어 많이 사용되고있다. 본 연구에서는 귀납학습을 각 질의어들에 대한 가장 적합한 문헌들을 추출하기 위해 사용하고 있다. 예를 들어 이들 귀납학습의 과정을 간단한 예를 통해 살펴보면 다음과 같다. 먼저 어떤 이용자가 신경망이 무엇인지 그리고 신경망에는 어떤 것이 있는지에 대한 HTML문서를 WWW상에서 검색하고자 한다고 했을 때 그 질의어로써 발췌할 수 있는 용어가 "신경망"이 될 수 있다. 그 다음 그 신경망에 관련된 가장 적합한 문서타이틀이 다음과 같다고 가정하자.

a. 신경망
b. 역전파 신경망
c. 인공신경망
d. 비감시(unsupervised) 학습(learning)
e. 오류전파에 의한 학습

이상의 문서들로부터 먼저 stop words를 제외한 용어들을 색인어로 추출할 수 있다. 그 용어들은 다음과 같다.

a. 신경망
b. 역전파
c. 인공
d. 비감시
e. 학습
f. 오류전파

이상과 같이 색인어가 결정이 되면 귀납학습을 통해 질의어가 여기서는 "신경망"이 들어가면 이상의 문헌들을 검색할 수 있는 가장 중요한 색인어를 결정할 수 있다. 여기서는 신경망에 해당되는 적합한 문헌을 검색하기 위한 가장 핵심이 되는 색인어 "신경망"과 "학습"이 귀납학습 프로그램실행 결과 선택되었다. 따라서 이상의 문헌들중에서 "신경망"이나 "학습"이 들어가 있으면 그 문헌은 "신

경망"에 관한 질문에 대한 적합한 문헌으로 검색이 되는 것이다. 이와 같이 귀납학습은 어떤 질의어에 대해 적합한 문헌을 선택할 수 있도록 그 질의어에 대한 가장 적합한 색인어들을 기계학습을 통해 결정해준다. 〈표 3〉 및 〈표 4〉는 본 논문에서 제안한 지능형 정보검색 시스템을 개발하기 위해 제시한 10개의 질의어 및 35개의 문헌타이틀에 대한 귀납학습을 적용한 결과인 의사결정 트리(decision tree)와 각 질의에 대한 핵심 색인어(discriminant descriptor)다. 〈표 3〉 및 〈표 4〉에서 보듯이 35개의 문서타이틀로부터 추출된 75개의 색인어중에 21개가 핵심 색인어로 추출되었다. 예를 들면 질의어 1(여기서 질의어 색인어는 composition,space, operator임)에 대한 핵심 문서 색인어는 "composition"이 추출되었다. 이런 결과는 종래의 검색시스템에서는 각 질의어에 대한 색인어를 문헌과 비교해서 문헌에 질의어가 포함되어있으면 적합한 문서로 검색되고 있다. 이런 결과, 질의어가 표현이 상이하거나 색인어간의 일관성이 없을 경우 검색효율성을 저하시킬 수 있는 것이다. 또한, 질의어 3(질의어는 "complex, analysis"임)에 대한 핵심 색인어는 귀납학습 실행결과 "complex, conformal"로 나타났다. 이런 결과는 비록 질의어에 "conformal"이라는 색인어가 포함되어 있지 않더라도 핵심색인어로 추출된 것이다. 따라서, 이러한 핵심 색인어를 추출함으로써 질의어 표현이 부적절하더라도 시스템이 추론해서 핵심 색인어를 추출할 수 가 있는 것이다. 따라서, 문서타이틀의 모든 키워드를 적합 문서 후보로 검색될 필요가 없고 단지 그 질의어 대한 핵심 색인어를 선정한

〈표 3〉 의사결정 트리

```
composition <=0: non-q1
composition > 0 : query1

graph > 0 : query2
graph <= :
    advanced <= 0: non-q2
    advanced > 0: query2

complex > 0: query3
complex <= 0:
    conformal <= :non-q3
    conformal > 0 : query3

certain >0: query4
certain <= 0:
    euclidean > 0: query4
    euclidean <=0:
        series <=: non-q4
        series > 0: query4

real > 0: query5
real <=0:
    measure > 0: query5
    measure <=0:
    contributions <=0: non-q5
    contributions > 0: query5

probability <=0: non-q6
probability > 0: query6

basis > 0:query7
basis <=0
    functional <=0: non-q7
    functional > 0: query7

mathematical > 0:query8
mathematical <=0:
    numerical <=0:non-q8
    numerical > 0: query8

average > 0: query9
average <=0:
    convolution > 0: query9
    convolution <=0:
        differential <=0: non-q9
        differential > 0: query9

rational > 0: query10
rational <=0
    representation <=0: non-q10
    representation > 0: query10
```

후 그 색인어를 포함할 경우에만 검색되므로 검색효율성이 매우 높을 것으로 기대되며 색인어가 그 표현에 있어 다르더라도 원하는 문서를 검색할 수 있을 것으로 보인다.

〈표 4〉 질의어에 대한 문서의 핵심색인어 추출결과

질의어	질의어	귀납학습적용결과 핵심 색인어
1	composition space operator	composition
2	graph theory combinatorial algorithm	graph combinatorics
3	complex analysis	complex conformal
4	series analysis fourier	certain euclidean series
5	real analysis elementary contributions	real measure contributions
6	probability process mathematical	probability
7	functional analysis multivariate	basis functional
8	numerical analysis methods modeling	mathematical numerical
9	convex theory differential equation analysis	average convolution differential
10	representation theory methods rational	rational representation

4.3 질의어와 문서 색인어 매핑을 위한 신경망 적용

일단 이용자의 질의어와 그에 대한 문서를 표현하는 색인어가 귀납학습 알고리즘을 통해 결정되고 나면 신경망을 이용해서 질의어와 문서 색인어를 연결시킨다. 신경망은 이용자 질의어가 다소 부정확하거나 불확실하더라도 현재 나타난 질의어와 과거에 학습한 내용을 바탕으로 적합한 문헌을 검색하는 특성이 있다. 따라서 이런 특성은 현재 인터넷을 이용하는 사람들의 질의어가 다양하고 다소 애매 모호하더라도 적합한 문서를 제공할 수 있을 것이다. 현재 지능형 정보검색을 위해 신경망을 설계하는데 가장 많이 사용되고 있는 네트웍은 역전파 신경망(back-propagation)이므로 본 연구의 신경망은 역전파 신경망을 이용하고 있다. 역전파 신경망에 대한 이론적 내용 및 관련분야는 김성희(1995)연구에 잘 나타나 있다. 일반적으로 역전파 신경망은질의어와 귀납학습 프로그램 실행결과로 얻어진 핵심 색인어를 기초로 해서 입력층(input layer), 출력층(output layer), 은닉층(hidden layer), 그리고 학습매개변수(learning parameter)로 구성된다. 구체적인 신경망 설계 및 구현 방법은 다음과 같다.

(1) 입 · 출력 및 은닉층내의 노드(node)수 결정

신경망은 일반적으로 입력(질의어), 출력(문헌), 은닉층으로 구분해서 설계하는데 여기서 입.출력의 노드수는 질의어(27개)와 문서 핵심어(21개)가 고정되어 있으므로 문제가 될 수 없지만 은닉층의 노드수는 문제가 되고 있다. 이미 알려져 있지만 은닉층의 노드수도 시스템 성능에 매우 큰 영향을 미친다. 은닉층 크기를 결정하는것과 관련해서는 통일된 방법이 없으므로 본 연구에서는 hill-climbing method를

〈표 5〉 1개의 은닉층에 대한 신경망시스템 성능

노드수	에 러	반복횟수
27	0.009	70
24	0.013	200+
25	0.009	90
26	0.009	81

〈표 6〉 2개의 은닉층에 대한 신경망 시스템 성능

은니층 1	은닉층2	에러	반복횟수
27	27	0.025	200+
27	24	0.026	200+
27	25	0.05	200+
27	26	0.04	200+
27	23	0.04	200+
27	22	0.34	200+
27	21	0.35	200+

사용하였다. 먼저, 27개의 입력층노드와 21개의 출력층노드로 트레이닝을 시작하여 다양한 수치를 입력해서 효율성을 측정하였다. 그 결과 〈표 5〉 및 〈표 6〉에서 보는 바와같이 은닉층을 하나추가하고 노드수를 27개로 하였을 때가 가장 빨리 converge하였다. 이런 결과는 은닉층을 하나만 추가하였을 때 이상의 데이터 셋트에서는 충분히 신경망이 수행되었다는 것을 의미한다.

(2) 학습매개변수 결정

역전파 신경망(back-propagation neural network)시스팀은 두 개의 매개변수 즉 모멘텀(momentum)과 학습율(learning rate)에 의해 강하게 영향을 받는다. 이 두개의 변수에 대한 이상적인 값을 결정하는 것도 본 연구의 주요내용이다. 이들 두 학습매개변수에 대한 이상적인 값도 아직까지 연구가 이루어지지 않았기 때문에 본 연구에서는 정보검색 시스팀에 신경망이 응용될 때 어떤 값이 가장 이상적인지에 대해 경험적인 방법을 사용해서 실험하였다. 여기서 경험적인 방법이라고 하는 것은 다양한 값 (대체로 0.0 - 0.9)을 적용해 봄으로써 어떤 값이 가장 적절한 지를 경험에 의해 결정하는 것이다. 이들 각 매개변수에 따른 성능결과는 〈표 7〉과 같다.

이상의 결과를 보면 일반적으로 학습률과 모멘텀이 증가할수록 converge속도는 빨랐다. 즉, 실험결과에 따르면 모멘텀이 0.9이고 학습률이 0.5일 때 converge가 가장 빨랐으며 다음은 모멘텀이 0.9이고 학습률이 0.5일 때 빨랐으며 모멘텀이 0.9이고 학습률이 0.9일 때는 다소 늦게 실행되어 모멘텀과 학습률이(0.9, 0.5) 또는 (0.5, 0.9)의 값을 갖는 것이 바람직한 것으로 보인다. 이러한 결과는 기존의 여러 연구(Cherkassky and Vassilas 1989; Gersho and Reiter 199)결과와 유사하게 나타났다. 예를 들면, Gersho and Reiter(1990)는 역전파 신경망을 이용한 다층 신경망 네트워크를 정보검색시스템에 응용하였는데 여기서는 학습률 0.6, 모멘텀을 0.9 사용하였다. 따라서, 학습시간을 최소화시키고 네트워크를 성공적으로 학습시키기 위해서는 학습률 및 모멘텀의 값을 시스템 설계시 고려해야할 것으로 보인다.

4.4 기계학습알고리즘을 이용한 지능형 정보검색 시스템의 검색효율성 측정

〈표 7〉 모멘텀과 학습율에 대한 성능비교

학습율	모멘텀	에러	반복횟수
0.5	0.9	0.01	24
0.9	0.9	0.01	76
0.5	0.5	0.01	80
0.9	0.5	0.01	34
0.5	0.25	0.01	120
0.9	0.25	0.01	65
0.5	0.1	0.01	146
0.9	0.1	0.01	78

기계학습방법중의 하나인 귀납학습과 신경망을 이용해서 지능형 정보검색 시스템을 구축한 후 이 시스템의 검색효율성을 테스트하였다. 이때 질의어는 위에서 제시한 10개의 질의어를 입력한 결과 100% 모두 정확한 적합문서를 검색하였다. 또한, 각 질의어를 구성하는 색인어 중에 임의로 1개의 색인어씩을 삭제해서 불완전한 질의어를 작성한 후 검색측정을 하였다. 그 결과 모든 검색결과가 완전한 질의어를 입력했을 때와 동일한 결과를 보여줬다. 이런 결과는 신경망 자체가 불완전한 입력을 하였을 때에는 추론능력이 있어 원하는 결과를 얻을 수 있는 특성 때문인 것으로 보인다. 즉, 신경망은 인간의 뇌 그리고 신경세포가 반응하는것과 유사하게 설계된 회로이다. 이는 많은 수의 소자를 네트워크로 연결하고, 각 소자들 사이의 연결의 세기(connection strenth)로 정보를 표현하고 기억한다. 따라서 이는 인간두뇌와 같이 비결정적인 특성을 가지고 있으므로 약간 틀리거나 비슷한 입력을 인식할 수 있기 때문인 것으로 분석된다.

5 결론 및 제언

본 연구에서는 웹상의 지능형 정보검색을 위해 기계학습알고리즘 적용에 대한 연구를 수행하였다. 본 연구의 결과를 요약하면 다음과 같다.

첫째, 웹상의 수많은 웹문서들간의 일관성 없는 색인문제점을 해결하기 위해 본 연구는 귀납학습(inductive learning algorithm)을 적용하였다. 귀납학습은 이미 알려진 질의어와 각 질의어에 대한 적합한 문헌을 사전에 학습시킴으로써 특정한 질의어에 대한 가장 중요한 문서 색인어가 어떤 것인지를 구별할 수 있게 된다. 또한 귀납학습은 이러한 각 질의어에 대한 적합한 색인어를 하나만 추출하는 것이 아니라 다양한 색인어들을 그 중요도와 함께 순위를 매길 수 있게 제공한다. 이런 결과가 의미하는 것은 앞으로 귀납학습알고리즘을 적용할 경우 문헌을 표현하는 용어들이 일관성이 없더라도 그 특정문헌을 표현하는 용어들에 대한 상대적인 중요도를 결정할 수 있다. 따라서, 검색효율성을 높일 수 있을 것이다. 문제는 얼마나 많은 색인어를 핵심색인어로 선정할 것인가이다. 즉, 재현률을 높이기 위해서는 가능한 많은 색인어를 핵심색인어로 선택해야 할 것이고 정도률을 높이기 위해서는 소수의 색인어를 핵심색인어로 의사결정트리로부터 선정해야 한다. 따라서 이것은 이용자의 선호도 및 검색시스템의 주제영역에 따라 달라져야 할 것이다. 즉, 이용자가 망라적인 검색을 원한다면 의사결정트리는 절단하지 말고 모두 핵심색인어로 선정해야 할 것이다. 그러나 그 반대일 경우에는 의사결정트리는 단순화시켜야 한다. 결국 귀납학습시스템은 상황에 따라서 다른 정보검색을 할 수 있는 융통성을 갖고 있다. 그러므로, 이런 핵심 색인어 수를 결정하는 연구가 검색효율성과 어떤 관계가 있는지 구체적인 연구가 이루어 져야 할 것이다.

둘째, 신경망을 이용함으로써 이용자의 질문이 다소 부정확하거나 불완전하더라도 신경망의 특성으로 인해 정확한 정보를 검색할 수 있다. 이는 인간두뇌와 같이 비결정적인 특성을 가지고 있으므로 약간 틀리거나 비슷한 입력을 인식할 수 있기 때문인 것으로 분석된다. 그러나, 본 연구에서는 단지 10개의 질의어 및 35개의 문서만으로 테스트하였기 때문에 다양한

주제분야와 많은수의 질의어 및 문서를 갖고 연구할 필요성이 있다.

셋째, 신경망설계시 시스템성능에 영향을 미치는 요소들을 경험적인 실험을 통해 검증하였다. 역전파 신경망은 일반적으로 입력(질의어), 출력(각 질의어에 대한 적합한 색인어), 그리고 입력층과 출력층사이에 있는 은닉층(hidden layer)로 구성이 되는데 이러한 요인들은 모두 시스템성능에 영향을 미치는 것으로 나타났다. 이런 결과는 Cherkassky and Vassilas(1989)의 연구결과와도 비슷한 결과를 나타냄으로써 효율적인 신경망 시스템을 설계하기 위해서는 이상의 요인들을 고려해야 할 것으로 보인다.

참 고 문 헌

김성희. 1999. The Application of Machine Learning Techniques for Intelligent InformationRetrieval. 『정보과학연구』, 동덕여자대학교 정보과학연구소, 2:145-168

김성희. 1996. The Study on the Effectiveness of Information Retrieval in the Vector Space Model and the Neural Network Inductive Learning Model. 『데이타베이스 저널』, 한국데이타베이스학회, 3(2): 75-96

오종인, 백준호, 최준혁, 이정현. 1997. 상호정보량을 이용한 색인어 분류에 의한 웹정보검색시스템의 정확도 향상. 『한국정보과학회 가을 학술발표 논문집』, 23(2): 201-204

조강래, 김형근, 신본기, 김영환. 1997. WWW에서 모니터링 에이전트. 『HCI's 97학술대회 발표논문집』, 한국통신 멀티미디어연구실.

조영재, 이창훈, 박태순. 1997. WWW상의 자료검색을 위한 효과적인 에이전트 검색알고리즘 구현. 『한국정보과학회 가을 학술발표논문집』, 24(2):77-80

Cherkassky, V. and N. Vassilas. 1989. "Performance of Back Propagation Networks forAssociative Database Retrieval." *Proc. International Joint Conference on Neural Networks*, Volume 1.

Courtois M. P. 1996. "Cool Tools for Web Searching: An Update." *Online* 20(3): 29-31

Courtois M. P., et al., 1995. "Cool Tools for Searching the Web: A Performance Evaluation." *Online* 19(6): 14-32

Doszkocs, T.E., J. Regiia, and X. Lin. 1990. *Connectionist Models and Information Retrieval, Annual Review of Information Science and Technology*. Vol.25.

Gershe, M. and R. Reiter. 1990. "Information Retrieval using Hybrid Multi-layer Neural Networks." *International Joint Conference on Neural Networks*. vol. II

Goldberg, D. and D. Nichols, B. M. Oki, and D. Terry. 1992. "Using Collaborative

Filtering weave an information Tapestry." *Communications of the ACM*, 35(12)

Mita, V. and T. Gedeon. 1991. "Automatic Text Analysis and Information Retrieval in Law Using a Neural Network." *Proceedings of thd 11th BCS IRSG Research Colloquium on Information Retrieval*: 188-189

Quinlan, J. R. 1988. "Decision Trees and Multi-Valued Attributes." *Machine Learning*, 11: 305-318.

Quinlan, J. R. 1986. "Induction of Decision Trees." *Machine Learning*, 1: 82-106.

Quinlan, J. R. 1992. C4.5: *Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Shaw, M.J., J. A. Gentry, and S. Piramuthu. 1990. " Inductive learning Methods for Knowledge-Based Decision Support: A Comparative Analysis." *Computer Science in Economics and Management*, 3: 147-165.

Sheth, B. and P. Maes. 1993. "Evolving Agents for Personalized Information Filtering." *In: Proceedings of the Ninth Conference on Artificial Intelligence for Application*. IEEE Computer Society Press.

Watters, C. R. 1989. "Logic framework for information retrieval." *Journal of The American Society for Information Science*, 40: 311-324.

Wilkinson, R. and P. Hingston. 1992. "Incorporating The Vector Space Model in a Neural Network used for Document Retrieval." *Library HI Tech*, 10: 69-75.

Wilkinson, R. and P. Hingston. 1991. "Using the Cosine Measure in a Neural Network for Document Retrieval." *14th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*: 202-210.

Yan, T.W. and H. Garcia-Molina. 1995. "A Tool for Wide-Area Information Dissemination." *In Proceedings of the USENIX 1995 Winter Technical Conference*. New Orleans, La. Jan.

〔부록 1〕 실험 질의어 및 그에 대한 적합한 문서

질의어 1:Composition Operators on the Spaces
적합문서
- Angular derivatives and compact composition operators on hardy and bergman spaces
- Composition operators
- The essential norm of a composition operator

질의어 2: Combinatorics
적합문서
- Graphs and hypergraphs
- Advanced Combinatorics
- Graph Theory

질의어 3: Complex Analysis
적합문서
- An introduction to complex analysis
- Complex Analysis
- Complex Analysis and its application
- Analytic functions of one complex variable
- Function theory of several complex variables
- Conformal mapping
- complex numbers and conformal mapping

질의어 4: Weak type estimates for maximal operators on certain Hp classes
적합문서
- A maximal function with applications to Fourier Series
- The molecular Characterization of certain hardy spaces
- Introduction to fourier analysis on Euclidean spaces

질의어5: Real Analysis
적합문서
- Concepts of real analysis
- Geometric measure theory
- Measure theory and integration
- Elementary real analysis
- A first course in real analysis
- Contributions to Fourier analysis

질의어 6: Stochastic process
- Introduction to Probability
- An introduction to probability and stochastic process

질의어 7: Approximation on Scattered Data
적합문서
- Radial basis function approximation: from gridded centers to scattered centers
- Multivariate interpolation with radial basis functions
- Function analysis

질의어 8: Numerical Analysis
적합문서
- Numerical Analysis: an Introduction
- Numerical Analysis and mathematical modeling
- Mathematical analysis
- Mathematical analysis and numerical methods

질의어9: Fourier integrals in classical analysis
적합문서
- Inequalities for strongly singular convolution
- Average in the plane over convex curves and maximal operators
- Pseudo-differential operators

질의어 10:Lie Theory and Representation theory
적합문서
- Rational methods: Lie Algebra
- Representation Theory