한국지능정보시스템학회논문지 제12권 제2호 2006년 6월(pp.145~165)

인지 매핑을 이용한 정보 필터링 시스템*

김진화
서강대학교 경영학과
(Jinhwakim@sogang.ac.kr)

이승훈
서강대학교 경영학과 박사과정
(au4240@sogang.ac.kr)

변현수
서강대학교 경영학과 박사과정
(elbim@sogang.ac.kr)

정보 필터링 시스템은 사용자의 요구를 충족시킬 수 있도록 설계되어 있으나 변화가 심한 사용자의 정보 요구를 충족시키기에는 정확도 저하 등의 문제점이 있다. 본 연구에서는 인간의 뇌에서의 정보처리과정을 시뮬레이션하는 인지적 브레인 매핑의 정보 필터링 시스템을 제안한다. 특정 단어나 패턴에 기초하여 필터링하는 기존의 필터링 시스템과 비교할 때 제안하는 필터링 시스템은 키워드와 키워드간의 관계를 이용하여 필터링을 하는 시스템이다. 본 연구는 키워드와 키워드간의 관계를 이용하여 정보를 기록저장하고, 저장된 정보를 지도화하여 필터링에 응용하는 것이다. 키워드를 이용한 필터링 방법과 키워드간의 관계를 이용하여 필터링하는 방법을 통합하여 필터링을 실시하고 필터링 성능에 영향을 미치는 방법을 검증하기 위해 각 방법별로 가중치를 적용하여 최적의 결합가중치를 도출해낸다.

논문접수일 : 2006년 2월 게재확정일 : 2006년 6월 교신저자 : 김진화

1. 서론

현대 사회의 특징 중의 하나로 정보화 사회를 들 수 있다. 정보화 사회란 사회 전체가 정보가치의 창출에 주력하고 정보통신기술을 통하여 대량으로 정보의 생산, 처리, 유통이 가능하고 정보 산업이 중심이 되는 사회를 말한다. 정보통신기술의 비약적인 발전으로 과거에 비해 정보를 효과적으로 창출, 처리, 관리, 통제, 저장하고 이를 빠르고 쉽게 전달할 뿐 아니라 확대된 지적 능력들간의 소통과 연계를 통해 연결된 정보의 양과 가치를 무한하게 증폭할 수 있다. 그러나 현대인들은 정보 과부하로 인한 불편을 겪고 있다. 개인사용자와 기관, 단체들은 공적인 관계와 교환을 통해 좀 더 많은 정보원과 대량의 정보를 이용하는 것이 가능하며 많은 양의 연구보고서, 도서, 영화, TV 프로그램들까지 사용자들은 쉽게 접근할 수 있고 이용할 수 있다.

그러나 쉽게 접근 가능하고 이용할 수 있는 정보가 증가할수록 특히 현재와 같은 인터넷 환경에서 정보 과부하라는 새로운 문제가 제기되고 있다. 따라서 정보 관련 시스템은 정보 과부하의 문제를 해결하기 위해서 사용자가 필요로 하는 정보를 찾을 수 있어야 한다. 정보 관련 시스템은 사용자가 관련이 있는 정보를 찾고 관계가 없는 대량의 정보를 선별을 통해 제거하는 것을 지원해야 한다. 이

* 본 논문은 2005년도 한국학술진흥재단의 지원에 의하여 연구되었음.

러한 일련의 과정을 정보탐색이라 할 수 있다.

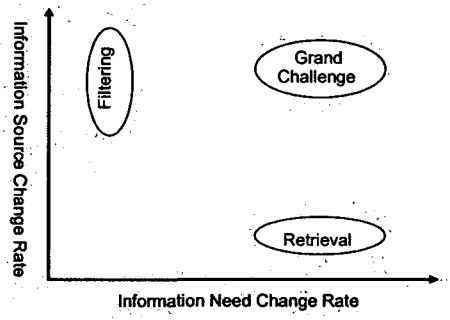
일반적으로 정보 탐색(information seeking)은 사용자의 목적으로부터 시작된다. 정보 탐색과 관련된 개념으로 정보 검색(information retrieval)과 정보 필터링(information filtering)이 있다. 정보 검색과 정보 필터링을 명확하게 구분하기가 쉽지는 않으나 <표 1>에서와 같이 정보 요구(information needs)와 정보원(information sources)의 특성으로 구분할 수 있다.

<표 1> 정보행위에 따른 정보요구와 정보원의 비교

구 분	정보 요구	정보원
정보 필터링	고정적, 특정적	유동적, 비구조적
정보 검색	유동적, 특정적	고정적, 비구조적
데이터베이스	유동적, 특정적	고정적, 구조적
정보추출	특정적	비구조적
속성적 정보제공	고정적, 특정적	유동적
브라우징	광범위함	불특정
엔터테인먼트	불특정	불특정

정보 필터링과 정보 검색은 매우 밀접한 분야로 둘 다 목표는 최종 사용자와 관련이 있는 정보를 찾아내어 지원하며 최종 사용자와 무관한 대량의 정보를 선별해 주는 것이지만 사용자 목적의 본질에 궁극적인 차이가 있다. 이러한 차이는 [그림 1]과 같이 설명할 수 있다. 정보 검색은 사용자의 관심이 변하는 것을 지원하기 위한 것으로 정보요구의 변화 속도는 높고 정보원의 변화는 낮다는 기본적 가정이 있다. 이와 달리 정보 필터링은 매우 유동적인 정보원에 대한 접근을 지원하는 것으로 이 경우 사용자의 관심은 비교적 고정적이라는 가정을 한다(Oard and Marchionini, 1996).

다양한 정보 이용 상황들은 이러한 가정사항들에 따라 만족하는 정보 검색, 정보 필터링을 이용하고 있다. 예를 들어 영화나 음악 분야에 대한 사용자의 관심은 고정적이므로 정보 필터링이 정보 이용문제에 있어서 적합하며 직무와 관련이 있는 응용 분야들에서는 정보 검색이 적합한 기술이다. 그러나 정보 요구의 변화와 정보원의 변화가 너무 빠르거나 경시하는 응용 분야에서는 이 두 가지 기술이 적합하지 않다. 이러한 분야는 특히 취급이 어렵기 때문에 정보 이용 시스템에서는 "대단한 도전"으로 간주하고 있다. 그 예로 주식시장의 경우와 같이 시장과 중개인의 관심이 끊임없이 변하는 분야를 들 수 있다.



[그림 1] Information Seeking Process

과거와 달리 현대 사회는 개인이 처리해야 하는 정보의 양도 기하급수적으로 증가하고 있다. 이러한 환경하에서 중요한 것은 필요한 정보를 신속하게 획득하는 것과 필요하지 않은 정보를 취하지 않는 것이다. 필요한 정보의 선별을 위해 정보기술을 이용하여 꼭 필요한 정보를 걸러내는 작업의 중요성이 증가하고 있다. 이와 같이 필요한 정보를 얻기 위해 정보를 선별하는 과정인 정보 필터링은 방대한 양의 정보가 중단 없이 생산되고 정보의 증가율이 높은 인터넷 분야에 활발히 적용되고 있다. 특히 유즈넷 뉴스, 전자우편과 같이 정

보의 증가율과 유입량이 많은 분야에 적용할 수 있다(최희윤, 1998). 기존의 정보 필터링은 이용자의 정보 요구에 따라 유동적인 정보원으로부터 적합한 정보를 선별하여 제공하는 것에 초점을 두어 정보 요구에 대한 사용자 프로파일의 구축시 고정적, 지속적인 정보 요구의 반영으로 인해 정보 선별에 있어서 정확성이 다소 저하될 수 있다. 따라서 본 연구에서는 정보 필터링 과정에 있어서 고정적인 사용자 프로파일로 인한 정확성의 저하를 제고시킬 수 있으며 인터넷과 같은 동적인 정보 이용 환경에 적합한 정보 필터링 시스템을 제안하고자 한다.

본 연구에서는 필터링 과정에 있어서의 정확성 제고와 동적인 정보 이용환경에 적합한 시스템을 구현하기 위해 가중 키워드 방식의 필터링 방법과 연결빈도 방식의 필터링 방법을 통합한 필터링 시스템을 제안하고자 한다. 특정 정보를 구성하고 있는 요소들의 빈도는 그 정보의 성격을 구분하는 데 중요한 역할을 한다. 또한 인접한 요소들은 상호 관련없이 독립적으로 존재하는 것이 아니라 인과관계로 연결되어 있어 필터링에 있어서 중요한 역할을 한다. 본 연구에서는 필터링을 통한 분류예측에 있어서 중요 역할을 수행하는 두 가지 방법을 통합한 후 적용된 두 방법에 가중치를 적용하여 필터링에 있어서의 기능상의 역할을 파악하였다.

본 연구에서 제안하는 필터링 방법은 인간이 정보를 획득하고 획득한 정보를 뇌에서 처리하는 과정을 시뮬레이션하는 것이다. 인간은 정보처리 과정에 있어서 시각 등의 감각기관으로 획득한 정보를 직렬적으로 처리하는 것이 아니라 개인이 학습하고 경험한 것에 기초하여 획득한 정보를 병렬적으로 처리하고 필요한 정보를 취사선택한다. 이러한 정보처리과정을 응용하게 되면 일괄적인 방법으로 필요/불필요의 기준을 수립하여 정보를 필터링하게 경우 필요한 정보를 차단하게 되는 문제를 해결할 수 있다. 이를 위해 인간의 정보처리 과정을 구체화하기 위해 정보의 빈도와 연결을 시각적으로 표현하는 인지매핑을 이용하였다.

2. 이론적 고찰

2.1 정보 필터링

네트워크 기술의 발달로 정보를 쉽게 공유할 수 있으며 이용할 수 있는 정보의 종류와 정보의 양이 증가하는 것과 병행하여 필요로 하는 정보를 찾기 위한 시간과 노력은 증가하게 되는 정보 과부하의 상황이 발생하고 있다. 정보 필터링은 인터넷의 발달과 더불어 나타나기 시작한 새로운 개념은 아니다. 주변에서 쉽게 정보 필터링의 예를 찾아 볼 수 있다. 특정 잡지를 구매하는 것은 그 잡지가 구매자와 무관한 내용을 담고 있는 것이 아니라 구매자가 관심과 관련이 있는 내용을 포함하고 있기 때문이다. 이러한 방법으로 정보이용자들은 이용하고 있는 정보에서 필터링을 수행하고 있다. 특정 잡지 내에서 이용자의 관심과 관련이 있는 정보를 담은 기사를 골라 낼 수 있듯이 이용자들은 획득한 정보에 대해서 계속적으로 필터링을 수행하고 있는 것이다. 그러나 정보통신 기술의 발전으로 대용량의 전자문서가 등장하게 됨에 따라 기존의 방법이 아닌 자동화된 방법으로 필터링하는 시스템이 필요하게 되었다(Foltz and Dumais, 1992).

이용자가 원하는 정보를 정보시스템으로부터 획득할 수 있는 모든 형태의 프로세스를 총괄하여 정보 탐색이라 정의한다면 정보 필터링은 인터넷의 성장과 함께 발전하기 시작한 네트워크 정보의 자동화된 전달에 따른 정보탐색의 한 분야로 정의

할 수 있으며 시간적으로 생성되는 대량의 정보를 이용자의 정보요구를 만족시킬 수 있도록 중요도에 따라 정렬하여 정보원을 제공하는 것을 의미한다(Oard, 1997). 또한 정보 필터링은 비정형 또는 반정형 데이터를 이용하는 시스템이며 대용량의 유동적인 문자정보를 다루는 시스템이라 할 수 있다(Belkin and Croft, 1992). 정보 필터링의 적용 영역으로 적합한 곳은 정보의 양이 방대하고 정보의 생성이 끊임없이 이루어지는 곳으로 즉 정보의 증가율이 높은 분야다. 현재 정보가 끊임없이 유입되고 생성되는 곳은 인터넷이다. 인터넷에서도 정보 필터링은 유즈넷 뉴스분야에서 시초를 이루어 전자우편과 웹에도 파급이 되었다. 정보 필터링이 필요한 분야는 정보가 기하급수적이며 시시각각으로 유입되고 원래부터 정보가 전자적으로 이루어져 시스템을 설계하고 실험할 수 있는 환경이 될 수 있으며(Oard, 1997) 정보의 양이 급속하게 증가될 것으로 예상되는 전자저널, 전자문헌과 영상, 멀티미디어분야까지 적용될 것이다. 이용자의 입장에서 정보의 형태보다는 내용을 중심으로 하여 적용영역을 살펴보면 일반적인 뉴스, 관심 분야의 서지정보, 학술저널 목차, 원문 정보 등으로 확장할 수 있다. 또한 이러한 모든 정보들이 인터넷을 통해 접근 가능한 디지털 정보로 변환되고 있어 향후 정보 필터링 영역은 방대한 정보를 디지털화 할 수 있는 모든 분야로 확장될 것이다.

정보 과부하의 문제를 해결하기 위해 사용하고 있는 정보 필터링 시스템은 사용자의 관심을 반영한 사용자 요구를 바탕으로 필터링을 수행하기 위해서는 많은 학습시간을 필요로 하며, 사용자 프로파일이 사용자별로 개인화되어 있지 않은 경우 효율적인 필터링을 기대하기 어렵다. 또한 새로운 사용자나 새로운 분야의 정보를 요구할 경우 사용자의 선호를 반영한 사용자 프로파일이 구축되어 있지 않을 경우 처음부터 학습을 새로 시작해야 하는 문제가 발생한다(양재영 등, 1999).

인터넷과 같은 대규모의 분산 네트워크 환경에서 이용자는 방대한 양의 정보를 검색하고 필요한 정보를 획득하기 위해 이용자가 직접 정보를 조합하고 가공해야 하는 어려움이 있다. 이를 지원하기 위한 정보 탐색엔진들이 있으나 중앙집중형의 색인방식으로 정보가 변경되었을 때 즉각적인 반영이 이루어지지 않으며 다양한 분류기준으로 인해 이용자가 정보를 찾는 것이 쉽지 않다. 이러한 불편을 해소하기 위해 출현된 것이 에이전트로 기존의 단순 반복적이면서 시간을 많이 소요하는 정보의 가공과 여과의 임무를 수행하는 소프트웨어라 할 수 있다.

정보 필터링 에이전트는 인터넷으로부터 이용자 프로파일을 구성하여 이용자의 관심이 있는 정보를 검색하여 이용자에게 제공해 주는 에이전트다(조영임, 2003). 정보필터링 에이전트가 정보검색의 부속물로 정의되기도 하는데 사용자의 정보요구로부터 저장된 추가 정보를 사용하여 불필요한 문서를 제거하고 검색된 문서 집합을 정제하기 때문이다. 정보필터링 에이전트는 웹 검색 에이전트와 같이 대량의 온라인 정보로 인한 정보 과부하의 문제를 다루지만 낮은 정확율의 문제점을 보완하기 위해 사용자 정보 프로파일을 다루는 점에서 차이가 있다.

웹 검색 에이전트가 사용자가 관심을 갖는 특정 웹 사이트를 찾는데 유용한 반면 정보 필터링 에이전트는 정보를 다양한 근원지로부터 모은 후 사용자 개인의 선호도에 기반하여 여과된 정보를 사용자에게 제공한다(Alper and Collin, 1997). 그러나 정보 필터링 에이전트 또한 사용자에게 제시하는 모든 텍스트를 다루는데 많은 시간과 노력이 필요하다는 문제점이 있다.

2.1.1 스팸메일 필터링(Spam mail Filtering)

스팸메일이란 발신자가 수신자의 동의 없이 전자메시지를 발송하거나 아무 관계가 없는 수신자에게 발송된 유익하지도, 원하지도 않은 전자메시지를 말한다. 스팸메일은 전자우편 이용자에게는 많은 시간과 비용을 낭비하게 하며 웹 메일 서비스 업자에게는 인터넷 체증 가중 및 통신 저하 등 유무형의 피해를 야기하고 있다. 최근 들어 바이러스 유포, 사용자 정보 해킹 등의 목적을 가진 스팸메일의 대량 유포로 사용자 정보를 해킹하거나 시스템에 바이러스를 감염시키는 등의 반 사회적인 문제점은 간과할 수 없게 되었다. 일반인들에게 이메일은 편지나 전화를 대체하는 도구로 활용되나 기업환경에 있어서 이메일은 업무 환경의 시간적, 공간적 문제를 극복할 수 있게 하며 사내 커뮤니케이션을 원활하게 하는 장점을 제공하며 회사 내 업무처리과정을 전체적으로 변화시키는 중요한 수단으로 성장하였다(신경식과 안수산, 2002).

그러나 스팸메일의 등장은 그 반대급부가 크게 나타나고 있다. 스팸메일은 개인에게는 정신적, 물리적 스트레스를 증가시키고 업무진행을 지연시키며 스팸메일 발송으로 메일 서버에 부하를 주게 되어 공공자원인 네트워크 자원을 독점하는 결과를 가져오고 있다(정재윤, 2000). 따라서 수신자에게 스팸메일이 도착하기 이전에 이메일을 필터링할 수 있는 방법에 대해서 많은 관심이 기울여지고 있다.

정옥란과 조동섭(2003)은 사용자의 메일 처리과정을 일정기간 관찰하여 각각 개인에 맞는 규칙을 형성하고 만들어진 규칙을 바탕으로 개인에게 불필요한 메일이나 스팸메일을 삭제토록 하는 개인화된 분류를 위한 웹 메일 필터링 방법을 제안하였으며 정확도를 높이기 위해 베이지안 알고리즘을 적용하였다.

신경식과 안수산(2002)은 데이터 마이닝 기법 중 분류 문제에 많이 사용되는 인공신경망과 의사결정나무 기법을 이용해서 스팸메일의 분류와 예측을 가능케 하는 모형을 구축하였다. 의사결정나무 기법을 적용한 스팸메일 필터링이 조금 더 나은 분류 성과를 보이고 있으나 영문 이메일 데이터를 이용한 점은 연구의 한계점이 될 수 있다.

서정우 등(2004)은 기존의 내용기반 이메일 필터링이 특정 단어의 패턴 매칭을 통해 필터링을 수행하나 스패머들이 제목이나 본문의 내용을 변형하여 스팸메일을 발송할 경우 이를 효과적으로 필터링하지 못하는 점에 착안하여 패턴 분류 문제에 특히 높은 성과를 보이는 SVM(Support Vector Machine)을 사용하여 생성된 색인어와 단어사전의 매칭을 통해 얻어진 데이터 셋을 SVM 분류기에 적용하여 정상 메일과 스팸메일을 분류하는 방법을 제시하였다. 특별히 많은 스팸메일의 내용을 차지하는 성인광고와 대출에 관련된 메일을 대상으로 실행한 결과 좋은 성능을 가지고 있음을 알 수 있었다.

조한철과 조근식(2002)은 스팸메일 필터링에 있어서 사용자가 직접 규칙을 작성할 필요 없이 학습을 통해 얻은 데이터로 자동으로 스팸메일을 필터링하는 시스템을 제안, 기존 문서분류에서 쓰이는 베이지안 학습방법 중 널리 쓰이는 통계적 알고리즘인 나이브 베이지안 분류자를 이용하였다. 기존의 규칙기반 시스템과 나이브 베이지안 분류자를 비교하여 분석한 결과, 오류율, 스팸재현율 등에 있어 나이브 베이지안 분류자를 이용하였을 때 정확도가 향상됨을 보이고 있다.

2.1.2 맥락 필터링(Contextual Filtering)

맥락 필터링의 활용분야는 매우 다양하다. 지문인식분야에서의 맥락 필터링은 인식된 지문 이미

지의 상세분류를 위한 지문 이미지 강화(Fingerprint Image Enhancement)에서 활용되고 있다. 맥락 필터링은 지문인식 분야에서는 낮은 품질의 지문 이미지를 인식하여 분류하는데 있어서 강력한 도구로 자리잡고 있다. 지문이미지에서는 지문 융기선의 시작과 융기선의 빈도, 융기선의 계속성, 경향 등의 정보를 스펙트럼 분석을 실시하여 지문 이미지를 성공적으로 분류하고 있다(Chikkerur et al., 2004).

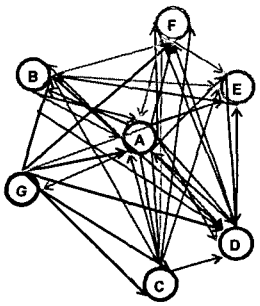
Uckun et al.(1999)은 일반 비행사고의 대부분은 비 계기비행 등급의 조종사가 의도적으로 계기비행을 하여야 하는 기상조건에 진입하였을 경우와 계기비행이 가능한 조종사가 천둥 번개, 우박, 돌풍과 같은 불안정한 기상환경에 처했을 때 발생하는 점에 착안하여 이를 해결하기 위해 AWARE라는 기상 요약 및 보고 시스템을 개발하였다. AWARE 시스템은 텍스트 기반의 정보와 그래픽한 비행 기상정보를 통합하여 임무상황과 장비특성에 맞는 상황경고를 제공하는 시스템이다. 최상의 비행 기상 브리핑은 조종사가 필요로 하는 기상정보를 원시 기상자료를 해석하여 제공하는 것이나 처리속도가 늦어 자동화된 서비스로 대체되었다. 그러나 자동화된 서비스는 방대한 양의 기상자료를 제공하나 계획된 비행임무 상황에 맞는 정보를 제공하지는 못하고 있다. 따라서 비행 기상 정보예보에 있어서 맥락 필터링은 상황이나 지역적 수준에 맞도록 조종사가 실제로 필요로 하는 비행경로상의 기상현상을 제공하고 있으며 이러한 방대한 양의 기상정보를 축약하여 제공할 수 있는 것은 맥락필터링으로 인해 가능하였다.

Ruch et al.(2001)은 의료보고서 편집의 향상을 목적으로 문장 내 존재하지 않는 단어로 인한 철자오류 정정 시스템을 개발하였다. 기존의 시스템들과 달리 의미론적이나 구문론적인 방법을 사용하였다.

2.2 연결빈도행렬(Connection Frequency Matrix)

연결 빈도 행렬은 인접 행렬의 개념에서 출발한다. 인접 행렬(Adjacency Matrix)은 데이터의 인접성(Adjacency)을 이용하여 의사결정공간(Decision Space)에서 유용하게 쓰일 수 있는 개념으로 품목 A와 품목 B가 존재할 때 품목 A와 B가 동시에 구매되었는지, 또는 품목 A가 B의 구매에 영향을 주었는지 그 여부를 확인할 수 있어 데이터 마이닝의 기법 중 연관규칙 분석(Berry and Linoff, 1997)이나, 또는 추천 시스템(Schafer et al., 2001) 및 데이터 시각화(Condon et al., 2002) 등에서 이용되고 있다.

이와 같은 데이터의 인접성은 유한개의 점과 선으로 구성된 도형인 연결그래프(Connected Graph)와 사상(Mapping)의 개념이나 도형의 위상적 성질을 이용하면 명확히 알 수 있다. 연결그래프에서 단위 정보를 표현하는 점을 "Vertex", 각 점을 잇는 선을 "Edge"라고 하고 또한 Edge에 방향성이 있는가에 따라서 유향 그래프(Directed Graph)와 무향 그래프(Undirected Graph)로 구분한다. 그 밖에 그래프 내에서 여러 Vertex들의 연결과정을 경로(Path)라고 하며, 시작점과 끝점이 연결된 경로는 특별히 순환(Cycle)이라고 한다. 연결그래프에서 선(Edge)이 점(Vertex)을 공유하고 있으면 1, 공유하고 있지 않으면 0으로 나타낸 행렬을 인접행렬이라고 한다. 인접행렬은 N개의 Vertex를 가지는 n * n 정방행렬이다. 인접행렬의 어떤 원소 Aij = 1이면 두 Vertex가 인접해 있다는 것이며, Aij = 0이면 두 Vertex는 인접해 있지 않은 것이다. [그림 2]는 연결그래프와 이에 대한 인접행렬의 예이다. 연결빈도행렬은 이러한 인접행렬의 특성에 방향과 누적빈도를 추가한 개념이다.



	A	B	C	D	E	F	G
A	0	2	3	2	1	2	1
B	2	0	1	2	2	1	3
C	3	2	0	1	2	2	1
D	2	3	2	0	1	1	2
E	1	2	3	2	0	2	1
F	2	1	2	1	1	0	1
G	1	2	2	2	2	2	0

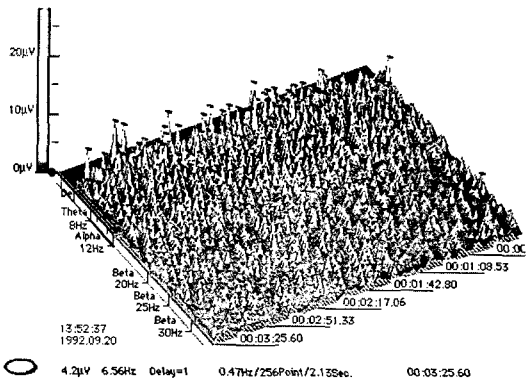
[그림 2] 연결그래프와 인접행렬의 예

2.3 브레인 매핑(Brain Mapping)과 인지 매핑(Cognitive Mapping)

인간의 뇌는 하나의 작은 우주라고 말할 수 있으며 인간의 뇌와 관련하여 연구할 분야는 무궁무진하다. 특히 인간의 뇌에서의 정보처리과정은 비중이 있는 연구분야이다. 인간의 뇌는 정보처리과정에 있어 주어진 프로그램에 따라 한번에 하나의 명령을 정보로 변환하고 이 정보에 기초하여 다음의 과정을 결정하고 한번에 하나의 정보를 처리하는 컴퓨터의 직렬 정보처리 방식과는 다르다. 인간의 뇌는 다수의 뉴런이 복잡하게 연결된 네트워크를 이루고 입력정보가 들어오면 다수의 뉴런에 전달이 되며 이러한 상호작용이 뇌의 전체에 퍼져 동시에 병렬적으로 정보를 처리하고 있다. 이러한 인간의 뇌에서의 정보처리 능력을 인공지능에 많은 응용을 하고 있다. 신경과학의 발달로 인해 뇌의 정보처리 능력을 인공지능에 응용하면서 뇌 연구의 지식영역이 점점 확대되고 있다. 인간의 뇌에 대한 연구는 기초과학, 공학, 의학, 심리학 등 여러 분야가 연관되어 있으며 미래형 핵심기술로 경제, 사회, 기술적 파급효과는 크다고 할 수 있다.

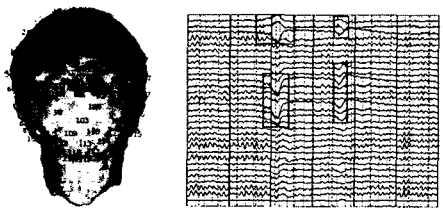
인간의 기능적 브레인 매핑(HFBM: Human Functional Brain Mapping) 분야는 크게 성장하고 있고 학제적 연구가 크게 집약되는 분야이다. 기능적 브레인 매핑의 개념은 의학에서 뇌의 활동을 시각화하기 위해 두피의 다른 지점 사이에서의 뇌의 전기 활동성을 측정하는 데에서 시작하였으며 측정의 결과는 뇌파도, 뇌전도라 하여 1928년 Hans Berger가 최초로 시도하였다(Law et al., 1991). 의학에서의 기능적 브레인 매핑에 대한 연구는 관련 학문들과의 학제적 연구를 통해 다양하게 연구되고 있다.

Fox et al.(2005)은 적절한 실험디자인을 통한 브레인 맵을 이용해 데이터 필터링에 응용할 수 있음을 보였으며 또한 브레인 맵 분류방법을 통해서 데이터베이스 구축을 위한 메타데이터 스키마를 만드는 데에도 유용함을 보였다. 또한 Law et al.(1991)은 전자 두뇌 그림(EEG: electroencephalogram)의 사용을 통해서 두뇌가 무엇을 하고 있는지를 시각화하는 방법을 확장 발전시켰는데 EEG는 인간의 뇌는 끊임없는 전기적 활동을 하고 있으며 이러한 활동의 결과는 기록할 수 있다는 것으로 기록은 뇌에 존재하는 수 천 개의 뉴런의 활동의 결과이다. 뉴런의 활동패턴은 인간의 심리 상태에 따라 달라지는데 빠르고 느린 뇌파의 패턴을 단순한 시각화로 나타내면 [그림 3]과 같은 패턴 특징을 나타낼 수 있다.



[그림 3] 3D Color EEG brain topography

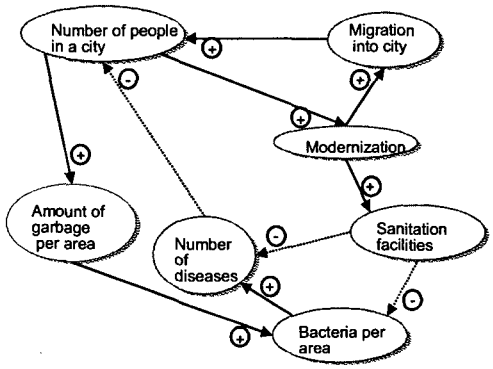
최근에는 [그림 4]와 같이 두뇌의 전기적 활동의 파라미터들을 EEG 지형도에 3차원과 칼라로 묘사하여 그릴 수 있는 소프트웨어의 개발로 두뇌의 전기적 활동을 3차원으로 재구성하여 나타내고 있다.



[그림 4] 뉴런의 활동 패턴과 뇌에서의 기억장소

인지매핑은 개인 또는 다수 사람들의 정신모형(Mental Model)을 그래픽으로 표현한 것이다. 인지매핑은 개념(Idea)과 이러한 개념들간의 연결(Link)로 구성되어 있으며 개념들의 사이의 연결들은 대부분 인과관계로 이루어져 있다(Rodhain, 1999). 이를 좀 더 구체적으로 표현하면 인지매핑이란 주어진 환경내의 요인들간에 존재하는 것으로 인지된 관계들의 표현이라고 할 수 있다. 인지매핑은 다양한 분야에 응용되어 다양한 문제들을 해결하고 있다. 예를 들어 정치분야에서는 정치인들이 지배력 향상에 따른 의사결정 개선을 돕기 위해 어떻게 의사결정을 하는가를 이해하기 위해 인지매핑을 이용하였다. 개인들이 내리는 의사결정은 정신모형에서 비롯되나 항상 그것을 인식하면서 의사결정을 내리지는 않는다. 그 이유는 의사결정에 필요한 지식이나 경험은 암묵적이며 형식적인 것이 아니기 때문이다(Axelrod, 1976). Eden et al.(1979)의 연구에서 인지매핑은 회사에서 이용하고 있는 도제 제도 과정에 대해서 잘 파악하지 못한 경영자들이 복잡한 조직내의 문제를 해결하기 위해 자신의 생각을 정리하고 조직화하기 위한 수단으로써 이용되었다.

인지매핑이 응용된 다른 분야로는 인사고과 분야이다. 캐나다 자동차 보험 회사에서는 직원 평가과정에 있어 모든 관리자가 각자가 선호하는 방식으로 평가를 했기 때문에 평가과정이 복잡하고 주관적이었다. 따라서 회사측에서는 모든 종업원을 공정하게 평가할 수 있는 시스템의 구축을 원했으나 이것은 모든 관리자가 각자의 개인적 평가과정을 모르고 있었기 때문에 어려운 문제였다. 따라서 관리자들이 그들의 부하직원을 평가하는 암묵적인 과정을 공식화하고 설명할 수 있는 수단으로 인지매핑이 이용되었다. 이와 같이 인지매핑이 응용되고 있는 분야는 이외에도 국제관계학, 경영학, 경영과학 등의 다양한 분야이다. [그림 5]는 공중 보건 분야에 대한 간단한 인지매핑의 예로써 +, -는 인과관계에 따른 긍정적, 부정적 영향 관계를 나타내고 있다(Montazemi and Conrath, 1986).



[그림 5] 공중보건분야에서의 인지매핑 예

3. 연구 방법

3.1 자료 수집

인지매핑을 이용하여 인간의 정보 처리과정을 시뮬레이션하는 필터링 시스템을 구현하기 위한

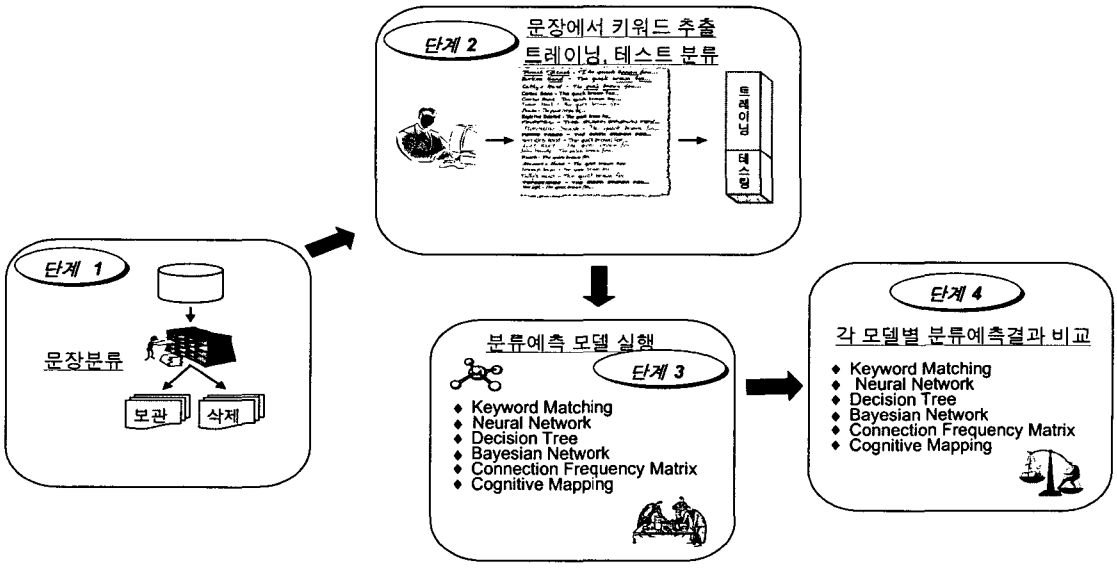
본 연구에서는 분석을 위해 피실험자 집단을 2개의 그룹으로 분류하였다. 그룹 1은 분석용 자료를 위해 문장을 작성하였으며 그룹 2는 작성된 문장을 분류하는 역할을 수행하였다. 그룹 1은 남녀 대학생 500명으로 구성, 문장을 작성하였다. 작성 문장은 100개의 단어를 이용하여 작성하였으며 사용된 단어는 국립국어연구원에서 2003년 5월에 발표한 한국어 학습용 어휘 목록 중 100개를 무작위로 선정하여 사용하였다. 피실험자는 1인당 5개의 문장을 주어진 단어를 이용하여 자유롭게 작성하도록 하였으며 한 문장당 주어진 100개의 단어 중 최소 3개 이상의 단어를 사용하여 문장을 작성하게 하였다.

수집된 총 2,500여 개의 문장 중 연구목적에 맞지 않는 문장을 제외한 2,300개의 문장을 획득하였다. 획득한 문장을 그룹 2의 250명의 대학생에게 배부하여 각각의 문장을 읽고 보관할 것인가 삭제할 것인가 결정하게 하여 "보관"과 "삭제"로 분류하게 하였다.

100개의 단어를 이용하여 문장을 작성한 이유는 신경망 분석의 경우 연구에서 처리하고자 하는 데이터에 포함되는 단어가 100개 이상이 되면 모형구축이 불가능한 점이 있어 사용단어의 수를 제한하였다. 이와 같은 실험결과는 베이지안 네트워크, 의사결정나무, 신경망, K-NN을 이용한 뉴스 기사 자동분류시스템 구축 연구(백용규와 서용무, 2003)에서도 신경망의 경우 실험에 포함된 단어 수가 증가할 경우 입력이 되지 않아 단어 100개 이상은 처리할 수 없어 다량의 단어를 이용한 실험은 베이지안 망과 의사결정나무를 이용한 것에서 확인할 수 있다.

3.2 분석절차

본 연구의 진행절차는 [그림 6]과 같은 절차에 따라 진행된다. 1단계에서는 분류 예측 분석을 위해 수집된 각 문장을 사전 처리하는 단계로 각 문장을 보관과 삭제로 분류한다. 2단계에서는 분류

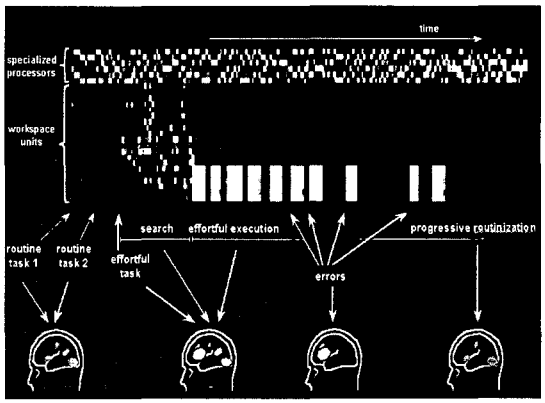


[그림 6] 분석절차

된 문장에서 키워드인 단어를 추출, 트레이닝용(2,000개)과 테스트용(300개)으로 분류하여 테스트 데이터가 트레이닝 데이터에 포함되지 않도록 한다. 3단계에서는 보관할 문장과 삭제할 문장으로 분류된 트레이닝 데이터를 이용, 학습한 후 테스트 데이터를 이용하여 기존의 분류 모델들과 본 연구에서 제안하는 인지매핑을 이용한 인지 필터링 시스템과의 분류 예측 정확도를 비교한다. 4단계에서는 3단계에서 실행된 모델들에서 산출된 결과를 이용, 각 모델의 예측정확도를 비교한다.

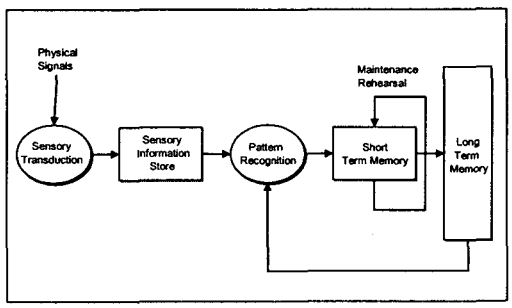
3.3 인지매핑(Cognitive Mapping)의 적용

본 연구에서 제안하는 인지매핑의 주요 개념은 인간의 정보처리과정을 시뮬레이션하는 것이다. 인간의 뇌는 약 수 십억 개의 신경세포와 이들을 상호 연결하는 약 수 십조 개의 시냅스로 구성되어 이들의 복합 작용에 의해 사람들은 사물을 인식하고 어떻게 행동할 것인지를 판단한다. 인지과학 분야에서는 이와 같은 인간의 두뇌에 의한 정신활동이나 신체기능을 추상적으로 다루지 않고, 구체적인 기술로서 재현하려고 한다. 즉 인간이 느끼고, 사고하고, 말로 표현하는 것을 추상적으로 표현하는 것이 아니라, 구체적 공식이나 절차로 재현하려고 한다. 인간의 뇌는 시각정보를 통하여 문자를 인식하고 의미를 이해한다. 인간의 뇌는 시각을 통해 획득한 정보들을 각각의 지정된 장소에 저장하며 저장된 정보는 지도와 같은 형태로 [그림 7]과 같이 나타낼 수 있다. 이러한 인간의 능력을 컴퓨터로 실현하려는 것이 패턴인식의 분야이며 이 분야에는 광학 문자 인식, 우편물 자동 분류, 문서인식, 도면인식 등의 분야가 부분적으로 실용화가 이루어졌으며, 최근에는 인공지능의 최신 기법인 신경망, 퍼지, 유전 알고리즘 등의 응용과 자연어처리, 심리학, 생리학, 인지과학 등 관련 학문과의 접목에 의해 문자인식 기술은 새로운 단계에 접어들고 있다.

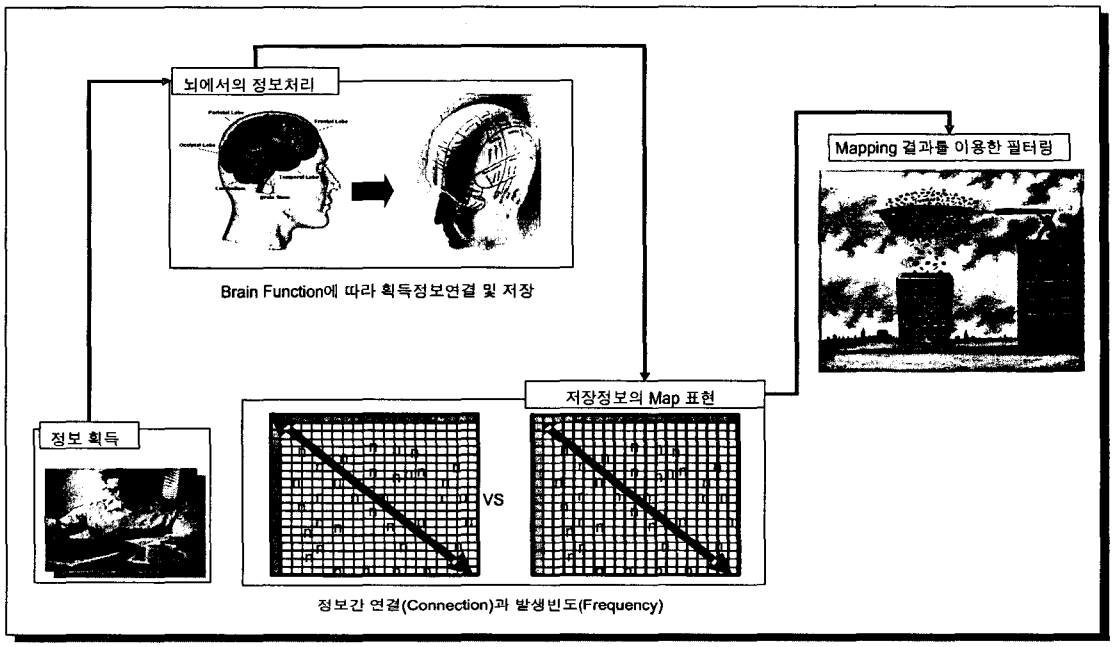


[그림 7] 뇌에서의 기억장소

인지과학에서는 인간을 정보처리기로 간주하고 있으며 정보처리행위는 인간의 사고체계의 기능과 행위간의 중요한 연결로 나타나고 있다. 인지과학에서 인간의 사고체계는 [그림 8]과 같이 감각기억, 단기기억, 장기기억으로 구성되어 있다. 이와 같은 구성은 정보처리, 정보 소비, 개념 획득, 추리, 의사결정 등을 포함하는 복잡한 기능들의 운영을 설명할 수 있는 정보처리와 관련된 구조이다(Liang, 1998).



[그림 8] 인간의 정보처리 모형



[그림 9] 인지매핑을 이용한 정보필터링 절차

본 연구에서는 문서에 사용한 단어의 출현 빈도뿐만 아니라 단어와 단어간의 연결은 새로운 의미를 창출하고 창출된 정보는 선택적 주의집중을 받으며 정보를 처리하는데 중요한 역할을 수행하게 된다는 정보처리 메커니즘을 필터링에 응용하였다. 예를 들어 필터링 대상 문장에 "여배우", "죽음"이라는 단어가 개별적으로 출현할 경우 이용자는 그 문장을 보관하기 보다는 삭제할 문장으로 분류할 확률이 높다. 그러나 유명 여배우의 죽음이라는 상황적 요인이 발생하였을 경우 이용자는 두 단어의 연결이 창출한 새로운 의미를 이용하여 정보처리과정에 활용할 것이다. 이와 같이 이용자가 살펴 본 단어들은 인간의 정보 인지과정에 따라 뇌의 특정 주소에 기록, 저장이 되고 저장된 정보는 개인이 축적해 온 경험과 학습에 따라 해당 단어들이 주는 정보의 가치를 판단될 것이다. 인간의 뇌에서 획득된 정보를 지도화하면 획득된 정보를 평면의 공간에 나타낼 수 있다. 획득된 정보의 지도화는 연결빈도행렬을 이용하여 구현하고 이를 이용하여 필터링에 적용한다.

이와 같이 단어의 인접성을 이용한 연결빈도행렬을 생성하기 위해 본 연구에서는 2,000개의 트레이닝 데이터를 보관과 삭제로 구분하고 출현한 단어에 각각 번호를 부여한 후 단어들간의 인접성을 이용하여 보관용과 삭제용의 연결빈도행렬을 생성하였다. 연결빈도행렬의 구현은 Microsoft사의 Visual Studio. Net의 Visual C++를 이용하여 구현하였다.

4. 실증분석

인지매핑을 이용한 필터링 시스템의 유용성을 검증하기 위한 비교대상으로는 문서분류시스템에

서 활용도가 높은 키워드매칭, 웨이티드 키워드매칭, 인공지능에서 성과가 높은 신경망, 의사결정나무, 확률적 학습방법을 채택하여 분석하는 베이지안 망, 단어의 인접성을 이용하는 연결빈도행렬을 이용하여 비교 · 분석하였다.

본 연구에서는 분류 예측의 정확도 측정에 있어서 타당성을 높이기 위해 총 10회에 걸친 교차검증을 실시하였다. 또한 문서분류에 있어서 단어의 출현 빈도와 단어간의 연결이 중요한 역할을 하는가를 파악하기 위해 인지매핑에 가중치를 적용하여 가장 높은 예측정확도를 보이는 가중치를 찾아내는 분석을 실시하였다.

4.1 의사결정나무 분석

문서 분류 예측을 위해 트레이닝용으로 분류된 데이터를 이용하여 의사결정나무 모형을 구축한 후 테스트용 데이터를 적용하여 분류 예측 정확도를 분석하였다. 테스트 데이터 300개를 이용하여 의사결정나무 분석을 실시한 결과, "보관"으로 예측한 93개의 사례 중 실제 "삭제"인 사례가 33개(35.5%)로 나타났고 "삭제"로 예측한 207개의 사례 중 실제 "보관"인 사례는 82개(39.6%)로 나타났다. 전체적인 의사결정나무의 분류 예측 정확도는 47.3%로 나타났다.

<표 2> 의사결정나무 분석결과

결 과		보 관(1)	삭 제(2)	합 계
보 관(1)	Count	60	33	93
	Row %	64.5	35.5	100
삭 제(2)	Count	125	82	207
	Row %	60.4	39.6	100
합 계		185	115	300

본 연구에서는 분석방법의 정당성을 확인하기 위한 방법으로 Cross Validation을 총 10회에 걸쳐 실시하였다. 10회에 걸쳐 실시한 교차검증 결과는 다음과 같다.

<표 3> 의사결정나무분석의 분류예측 정확도

방법	예측정확도(%)									
	1차	2차	3차	4차	5차	6차	7차	8차	9차	10차
의사결정나무분석	47.3	45.8	54.0	48.3	52.0	52.0	51.33	55.0	50.0	43.0

4.2 신경망분석

문서 분류 예측에 대한 신경망 구축과 평가를 위한 구축모형은 의사결정나무분석의 구축모형과 거의 동일하며 모형화 노드에 있어서 차이가 있다. 신경망 노드에서의 분석 방법으로 Clementine 8.1 프로그램에서의 디폴트인 Quick을 사용하였으며 트레이닝용으로 분류된 데이터를 이용하여 신경망 모형을 구축한 후 테스트용 데이터에 적용하여 분류 예측 정확도를 분석하였다.

<표 4> 신경망 분석 결과

결 과		보 관(1)	삭 제(2)	합 계
보 관(1)	Count	147	83	230
	Row %	49	27.7	100
삭 제(2)	Count	38	32	70
	Row %	12.7	10.7	100
합 계		185	115	300

테스트 데이터 3000개를 이용하여 신경망 분석을 실시한 결과, "보관"으로 예측한 230개의 사례 중 실제 "삭제"인 사례가 83 사례(28%)로 나타났고 "삭제"로 예측한 70개의 사례 중 실제 "보관"인 사례는 38 사례(12.7%)로 나타났다. 전체적인 신경망의 분류 예측 정확도는 59.7%로 나타났다. 본 연구에

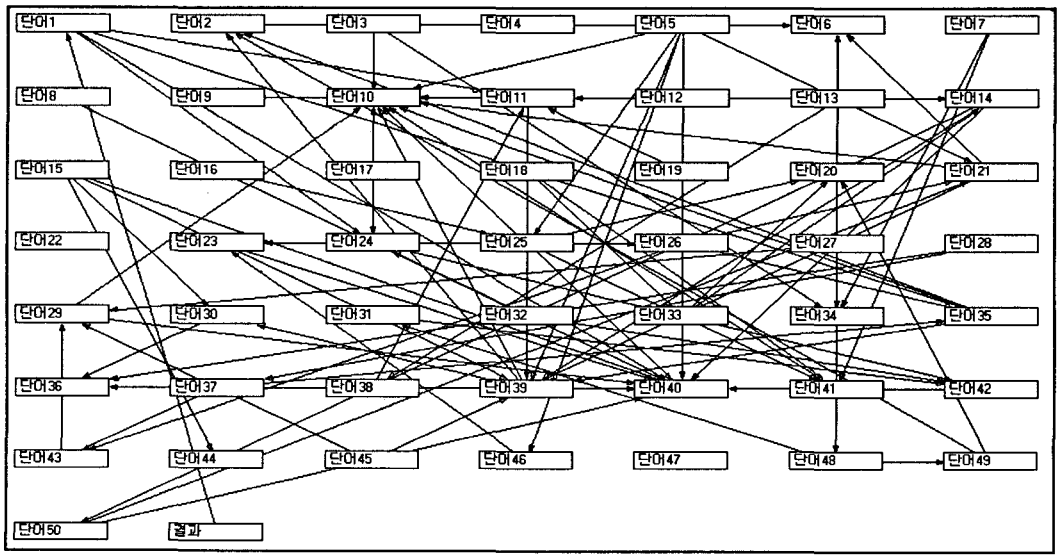
서는 분석방법의 정당성을 확인하기 위한 방법으로 교차검증을 총 10회에 걸쳐 실시하였다. 10회에 걸쳐 실시한 교차검증 결과는 다음과 같다.

<표 5 >신경망분석의 분류예측 정확도

방법	예측정확도(%)									
	1차	2차	3차	4차	5차	6차	7차	8차	9차	10차
신경망 분석	59.7	61.7	61	63.3	61.3	64.3	61	60	57	64

4.3 베이지안 망

베이지안 망은 많은 변수들간의 확률관계를 가시적으로 표현하는 모델이기는 하지만 변수의 개수가 많아지면 탐색공간이 넓어 학습하는데 시간이 많이 걸리기 때문에 탐색공간을 줄이기 위해 덜 중요한 특징들을 판단하여 축소하는 방법들을 사용해야 하는 제한이 있다(조한철과 조근식, 2002). 본 연구에서는 먼저 트레이닝 데이터를 이용하여 분류예측 모형을 구축한 후 구축된 모형의 결과를 테스트 데이터에 적용하여 분류 예측의 정확도를 측정하였다. 베이지안 망의 분류 예측 정확도를 계산하기 위해 베이지안 망 학습시 조건부 독립성 기반의 알고리즘을 사용하여 이를 구현한 Cheng의 BN Power Constructor 1.0 프로그램을 이용하였다. Cheng의 BN Constructor 1.0 프로그램은 모두 3개의 세부 모듈로 구성되어 있다, 첫 번째 모듈에서는 분석 대상 데이터의 사전처리 준비를 하는 Data Preprocessor 단계로 이 단계에서는 데이터 분석의 전처리를 위해 특징을 축소하거나 필요한 경우 데이터의 정제, 변환, 이산화의 작업이 이루어진다. 두 번째 모듈인 Power Constructor에서는 분석 대상 데이터의 사전 지식을 미리 입력하여 탐색공간을 줄일 수 있으며 사전 지식을 이용하여 실험대상인 변수들 간의 원인과 결과관계와 우선 순위를 정할 수 있다(Cheng and Greiner, 1999). 본 연구에서의 보관과 삭제여부를 표시하는 목표변수 "결과"와 변수들 간의 관계를 분석한 결과는 [그림 10]과 같다.



[그림 10] 목표변수와 설명변수간 관계도

트레이닝 데이터를 이용하여 구축된 베이지안 망 분류기의 학습 모형을 테스트용 데이터에 적용하여 산출한 분류 예측 정확도는 59.3%의 예측 정확도를 나타냈다.

<표 6> 베이지안 망의 Confusion Matrix

Predicted	보관(1.0)	삭제(2.0)	Lift Index
보관(1.0)	165	20	0.50036
삭제(2.0)	102	13	0.52278

베이지안 망 분석의 정당성을 확인하기 위해 총 10회에 걸쳐 실시한 교차검증결과는 다음과 같다.

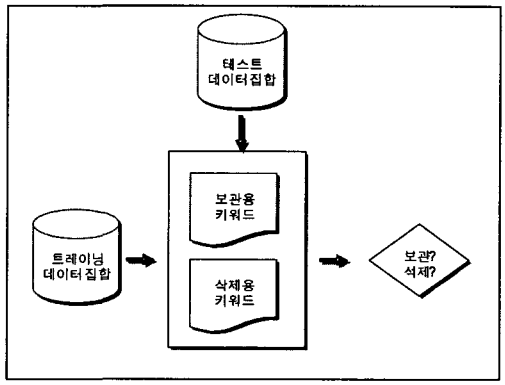
<표 7> 베이지안 망의 분류예측정확도

방법	예측정확도(%)									
	1차	2차	3차	4차	5차	6차	7차	8차	9차	10차
베이지안 망	59.3	64	59.3	64	61	61.7	61	55.7	59	65

4.4 웨이티드 키워드 매칭

문서분류에서 키워드 매칭방법은 해당 키워드의 카테고리 존재여부를 이용하여 분류 예측의 정확도를 측정하는 방법이나 그 정확도가 낮다. 웨이티드 키워드 매칭은 키워드 매칭 방식에 키워드 매칭의 낮은 분류 정확도를 높이기 위해 각 해당 카테고리에 출현하는 키워드의 빈도를 이용하여 분류하는 방법이다.

1차적으로 트레이닝 데이터를 이용, 보관용과 삭제용 카테고리를 구축한 후 테스트 자료를 이용하여 각각의 카테고리에 해당 데이터의 존재 유무를 확인한다. 해당 카테고리에 데이터가 존재하는 경우 카테고리별로 나타난 키워드의 빈도를 합산하는 방법이다. 위와 같은 방법으로 산출한 분류 예측 정확도는 61.7%의 예측 정확도를 나타냈다. 웨이티드 키워드 매칭 분석의 정당성을 확인하기 위해 총 10회에 걸쳐 실시한 교차검증결과는 다음과 같다.



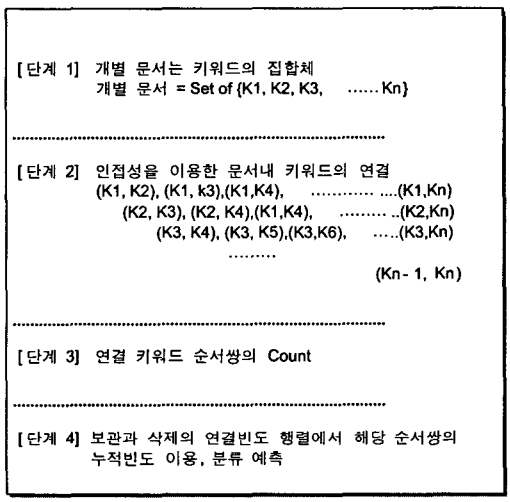
[그림 11] 웨이티드 키워드 매칭 분석 모형

<표 8> 웨이티드 키워드 매칭의 분류예측 정확도

방법	예측정확도(%)									
	1차	2차	3차	4차	5차	6차	7차	8차	9차	10차
웨이티드 키워드 매칭	61.7	64	59.3	64	61	61.7	61	55.7	59	65

4.5 연결빈도행렬

단어의 인접성을 이용하여 매트릭스를 만드는 개념인 연결빈도행렬을 이용하면 문서에서 사용된 키워드들로 생체인식의 한 분야인 지문인식과 같이 특정 패턴을 만들 수 있다. 본 연구에서 2,000개의 트레이닝 데이터를 보관과 삭제로 구분하고 출현한 단어에 각각 번호를 부여한 후 단어들간의 인접성을 이용하여 순서쌍으로 연결한 후 보관용과 삭제용의 연결빈도행렬을 생성하였다. 연결빈도행렬을 나타내는 알고리즘은 [그림 12]과 같이 요약할 수 있다.



[그림 12] 연결빈도 행렬 알고리즘 요약

트레이닝 데이터를 이용하여 각각 보관과 삭제 연결빈도행렬을 생성, 테스트 데이터를 보관/삭제로 구축된 각각의 연결빈도행렬과 조회하여 해당 행렬 값이 큰 쪽으로 분류한 결과 58%의 분류 예측정확도를 보였다. 연결빈도행렬 분석의 정당성을 확인하기 위해 총 10차에 걸쳐 실시한 교차검증 결과는 다음과 같다.

<표 9> 연결빈도행렬의 분류예측정확도

방법	예측정확도(%)									
	1차	2차	3차	4차	5차	6차	7차	8차	9차	10차
연결빈도 행렬	58	59	60.7	61	60.3	57	56.3	50.3	54	61.7

4.6 인지매핑

단어의 인접성을 이용하여 분류예측을 하는 기본적인 개념은 연결빈도행렬과 유사하나 인지매핑은 연결빈도행렬과 단어별 출현빈도를 누적하여 분류예측에 사용하는 방법인 웨이티드 키워드 매칭을 혼합한 형태의 분류예측 방법이다.

웨이티드 키워드 매칭에서 입증되었듯이 단어의 출현빈도는 분류하고자 하는 문서의 성격을 강하게 나타내고 있다. 또한 연결빈도행렬에서 단어간 인접성을 이용한 연결은 문서의 성격을 특정패턴으로 나타낼 수 있다. 앞서 전술한 바와 같이 지문인식에서의 지문의 특성을 나타내는 융기와 골의 역할을 웨이티드 키워드 매칭과 연결빈도행렬이 수행하는 것이다.

연결빈도행렬에 단어의 출현 빈도를 추가한 후, 테스트 데이터를 보관/삭제로 구축된 각각의 인지매핑과 조회하여 해당 행렬 값이 큰 쪽으로 분류한 결과 62%의 분류 예측정확도를 보였다. 인지매핑 분석의 정당성을 확인하기 위해 총 10차에 걸쳐 실시한 교차검증 결과는 다음과 같다.

<표 10> 인지매핑분석의 분류예측정확도

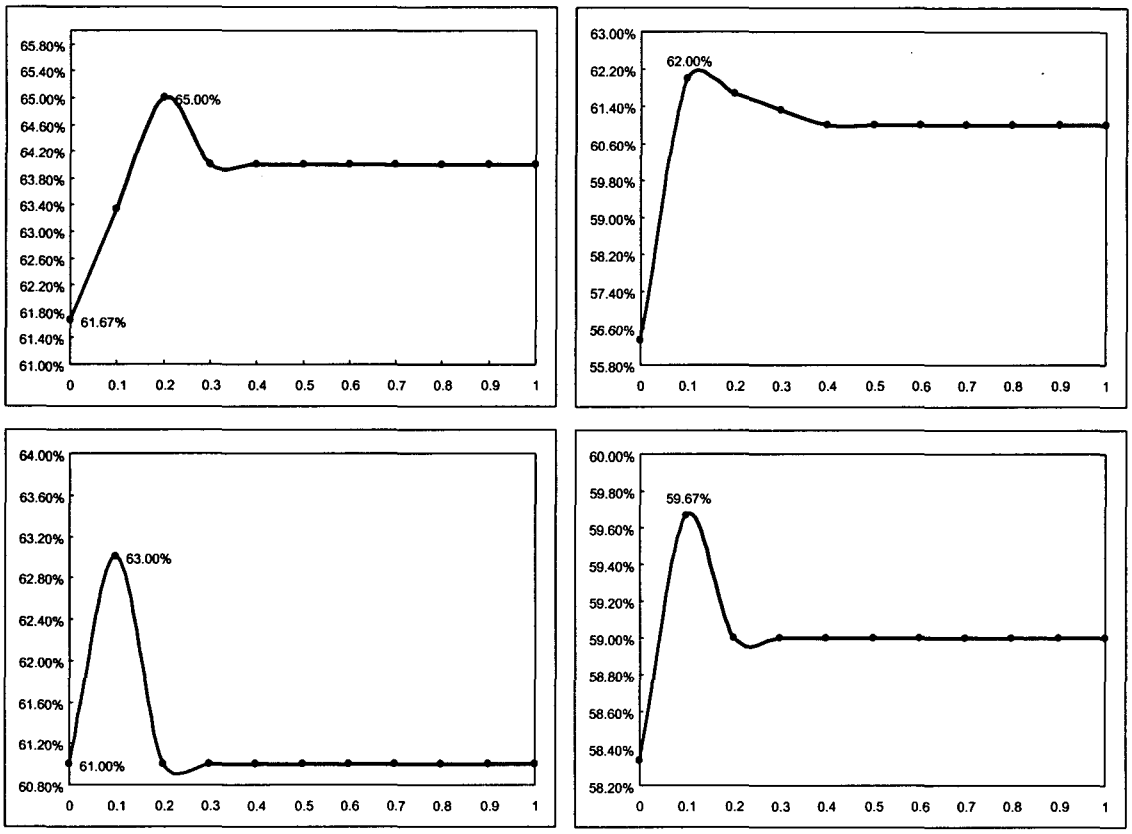
방법	예측정확도(%)									
	1차	2차	3차	4차	5차	6차	7차	8차	9차	10차
인지매핑	62	63.7	59.3	65.0	63.0	61.7	62.0	55.7	59.7	65.0

또한 인지매핑은 두 가지의 필터링 방법이 통합된 형태를 가지고 있어 이들 방법간의 유효성을 검증하기 위해 웨이티드 키워드 매칭방법과 연결빈도행렬 방법에 상대 가중치를 적용하여 분석하였다. 분석결과의 정당성을 확인하기 위해 총 10차에 걸쳐 교차검증을 실시하였다.

통합된 인지 매핑 필터링 방법의 역할을 파악하고자 가중치를 각 방법 별로 적용하여 분석한 결과 연결에 적용한 가중치가 0.1~0.3에서 예측정확도가 증가하고 그 이후의 값에서는 정확도의 변화가 없었다([그림 13] 참조). 단어의 빈도와 인접한 단어의 연결빈도가 필터링에 있어서 중요한 역할을 수행하며 그 중에서도 연결빈도의 역할이 크다는 것을 알 수 있다.

<표 11> 인지매핑 가중치 적용결과

적용가중치		예측 정확도(%)									
WK	CFM	CV1	CV2	CV3	CV4	CV5	CV6	CV7	CV8	CV9	CV10
1	0	65.7	61.7	62.0	61.7	61.0	58.7	56.3	53.7	58.3	62.7
0.1	0.9	62.0	62.3	59.0	63.3	63.0	61.7	62.0	54.7	59.7	63.7
0.2	0.8	61.7	62.7	59.3	65.0	61.0	61.7	61.7	55.7	59.0	63.3
0.3	0.7	61.7	63.7	59.3	64.0	61.0	61.7	61.3	55.7	59.0	65.0
0.4	0.6	61.7	63.7	59.3	64.0	61.0	61.7	61.0	55.7	59.0	65.0
0.5	0.5	61.7	63.7	59.3	64.0	61.0	61.7	61.0	55.7	59.0	65.0
0.6	0.4	61.7	63.7	59.3	64.0	61.0	61.7	61.0	55.7	59.0	65.0
0.7	0.3	61.7	63.7	59.3	64.0	61.0	61.7	61.0	55.7	59.0	65.0
0.8	0.2	61.7	63.7	59.3	64.0	61.0	61.7	61.0	55.7	59.0	65.0
0.9	0.1	61.7	63.7	59.3	64.0	61.0	61.7	61.0	55.7	59.0	65.0
0	1	61.7	63.7	59.3	64.0	61.0	61.7	61.0	55.7	59.0	65.0



[그림 13] 가중치 적용에 따른 예측정확도의 변화

<표 12> 분류예측정확도 분석결과 종합

(단위: %)

구분	웨이티드 키워드매칭	연결빈도행렬	*의사결정나무**	*신경망**	베이지안 망	인지매핑
1차	61.67	58.00	*47.33*	*59.67*	59.33	62.00
2차	64.00	59.00	*45.67*	*61.67*	64.00	63.70
3차	59.30	60.70	*54.00*	*61.00*	59.33	59.30
4차	64.00	61.00	*48.33*	*63.33*	64.00	65.00
5차	61.00	60.30	*52.00*	*61.33*	61.00	63.00
6차	61.70	57.00	*52.00*	*64.33*	61.67	61.70
7차	61.00	56.30	*51.33*	*61.00*	61.00	62.00
8차	55.70	50.30	*55.00*	*60.33*	55.67	55.70
9차	59.00	54.00	*50.00*	*57.00*	59.00	59.70
10차	65.00	61.70	*43.00*	*64.00*	65.00	65.00
평균	61.19	57.83	*49.87*	*61.37*	61.19	61.70

* 음영으로 처리한 부분은 의사결정나무분석과 신경망 분석결과로 입력변수의 크기가 100개를 초과할 경우 모형구축이 불가한 제한점이 있다(백용규, 서용무(2003)의 연구 참조).

4.7 분석결과 종합

인간의 정보처리과정을 시뮬레이션하는 인지매핑을 이용한 정보필터링 시스템의 성능을 비교하기 위해 기존 기법들과의 비교분석을 실시하였다. 분석결과의 정확성과 타당성을 위해 트레이닝과 테스트용으로 자료를 분리하여 분석하였으며 10회에 걸쳐 교차 검증을 실시하였다.

분석결과 높은 예측정확도를 보이는 분석 기법은 본 연구에서 제안하는 인지매핑기법, 신경망, 베이지안 네트워크, 웨이티드 키워드 매칭, 연결빈도행렬, 의사결정나무 순으로 분류예측의 정확도를 보이고 있다. 분류예측의 기법 중 하나인 키워드 매칭기법을 비교 분석기법으로 포함하여 분석하고자 하였으나 키워드 매칭의 경우 단어의 출현여부에 따라 분류하는 방식으로 키워드의 존재여부로만 분류하는 것은 예측정확도가 낮을 것이라는 예상과 같이 예측정확도는 매우 낮아 본 연구결과에서는 제외하였다.

5. 결론

정보통신 기술의 발전에 따라 인간이 처리해야 하는 정보는 전자 문서와 같은 형태적인 변화와 처리해야 하는 정보의 양이 증가하는 양적인 변화를 보이고 있다. 취급하는 정보의 증가는 정보 과부하라는 반대급부 또한 가지고 있어 증가하고 있는 정보에 대한 필터링은 점점 중요한 영역이 되어 가고 있으며 적용 범위 또한 확장되고 있다. 이와 같이 정보를 획득하여 필요한 정보만을 의사결정 또는 지식창출에 이용해야 하는 현대 사회에서 정보필터링은 중요 연구영역이라 할 수 있다. 본 연구에서 제안한 인지 매핑을 이용한 정보 필터링 시스템은 인간의 정보처리 과정을 필터링에 응용하였다는 의의가 있다. 특정 정보 시스템만을 이용하여 정보 필터링을 할 수 있는 것이 아니라 일상 생활에서 자신도 모르는 사이 정보 필터링을 수행하고 있는 것과 같이 인간의 뇌에서 정보를 획득하여 필요/불필요의 정보를 분류하는 것을 시

뮬레이션하여 정보 필터링에 응용하고자 한 본 연구는 기존의 분류모델들인 신경망, 의사결정나무, 베이지안 망, 키워드 매칭, 웨이티드 키워드 매칭, 연결빈도행렬 등과의 예측 정확도의 비교시에도 우수성이 입증되었으며 입력단어 수에 제한을 받지 않으므로 활용측면에서도 우수하다고 할 수 있다. 본 연구가 가지는 또 다른 의의는 특정 단어 또는 패턴만을 이용하여 필터링하는 기존 시스템과는 달리 단어의 존재, 단어와 단어를 연결, 활용함으로써 지식, 경험, 통찰력 등을 이용하여 획득한 정보를 처리하는 인간의 정보 처리행위를 시뮬레이션하였고 분류예측에 있어서 출현 단어 빈도의 역할을 검증하기 위해 가중치를 적용하여 분류예측에 활용한 점에서 의의가 있다.

그러나 이용자의 요구에 맞는 정보를 얻기 위해 사용하는 정보 필터링 시스템이 이용자의 의도와 다르게 정보를 분류하거나 이용자의 다양한 요구를 반영하지 못할 때는 정보 필터링을 사용하지 않은 경우 보다 못할 수 있다. 정보 과부하의 문제를 해결하기 위해 채택한 필터링 시스템이 불필요하게 복잡한 필터링 단계를 설정하거나 등급을 과도하게 설정할 경우 필요한 정보를 차단하게 되는 역효과가 있을 수 있다. 본 연구가 가지는 한계점으로는 특정 분야에 대한 인간의 관심을 중심으로 정보를 필터링하게 될 경우 시시각각으로 변화하는 관심에 대한 지속적인 갱신이 수반되지 않을 경우 생길 수 있는 필터링의 문제점을 해결하기 위해 인지매핑이라는 개념을 소개하고 그 성능을 확인하였으나 성능을 검증하기 위해 수집한 자료의 범위가 다소 제한적이었다. 제한적인 범위로 인해 생길 수 있는 결과 해석의 문제를 10회에 걸쳐 실시한 교차검증으로 보완하였다. 차후 연구에서는 본 연구에서 제안하는 개념을 특화되고 특정의 정보를 다루는 분야에 적용할 경우 객관적으로 성능을 확인할 수 있을 것이다.

참고문헌

[1] 백용규, 서용무, "인터넷 뉴스기사에 대한 자동 분류 정보 시스템에 관한 연구", *한국 경영정보학회 추계학술대회 발표논문집*, (2003), 574-581.

[2] 서정우, 손태식, 서정택, 문종섭, "n-Gram 색인화와 Support Vector Machine을 사용한 스팸메일 필터링에 관한 연구", *한국정보보호학회 정보보호학회논문지*, 14권 2호(2004), 23-34.

[3] 신경식, 안수산, "데이터마이닝 기법을 활용한 스펨메일 분류 및 예측 모형 구축에 관한 연구", *한국지능정보학회 학술대회 발표논문집*, 2권(2000), 359-366.

[4] 양재영, 홍광희, 최중민, "효율적 정보 필터링을 위한 지능형 협동 정보 필터링 에이전트", *한국정보과학회 학술발표논문집*, (1999), 69-71.

[5] 정재윤, *이메일 마케팅*, 웹 마니아, 2000.

[6] 정옥란, 조동섭, "개인화된 분류를 위한 웹메일 필터링 에이전트", *한국정보처리학회 정보처리학회논문지 B*, 10권 7호(2003), 853-863.

[7] 조영임, *인공지능시스템*, 홍릉과학출판사, 2003.

[8] 조한철, 조근식, "나이브 베이지안 분류자와 메시지 규칙을 이용한 스팸메일 필터링 시스템", *한국정보과학회 봄 학술발표논문집*, 29권 1호(2002), 223-225.

[9] 최희윤, "정보필터링을 이용한 주문형 정보서비스에 대한 연구", *정보관리학회지*, 15권 1호(1998), 63-81.

[10] Alper. K. C. and G. H. Collin, *Agent Sourcebook*, Wiley, New York, 1997.

[11] Axelrod, R., *Structure of Decision*, Princeton, NJ: Princeton University Press. 1976.

[12] Belkin, N. J. and W. B. Croft, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?", *Communication of ACM*, Vol.35(1992), 29-38.

[13] Berry, M. and G. Linoff, *Data Mining Techniques*, Wiley, New York, 1997.

[14] Cheng, J. and R. Greiner, "Comparing Bayesian Network Classifiers", *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence,* (1999), 101-107.

[15] Chikkerur, S., C. Wu, and V. Govindaraju, "Systematic approach for feature extraction in Fingerprint images", *1st International Conference on Biometric Authentication*, (2004), 344-350.

[16] Condon, E., B. Golden, S. Lele, S. Raghavan, and E. Wasil, "A visualization model based on adjacency data", *Decision Support Systems*, Vol.33, No.4(2002), 349-362.

[17] Eden, C., S. Jones, and D. Sims, *Thinking in Organisations*, Macmillan, London, 1979.

[18] Foltz, P. W. and S. T. Dumais, "Personalized Information Delivery: an Analysis of Information Filtering Methods", *Communications Of The ACM*, Vol.35, No.12(1992) 51-60.

[19] Fox, P. T., A. R. Laird, S. P. Fox, P. M. Fox, A .M. Uecker, M. Crank, S. F. Koenig, and J. L. Lancaster, "BrainMap Taxonomy of Experimental Design Description and Evaluation", *Human Brain Mapping*, Vol. 25(2005), 185-198.

[20] Liang, T. Y., "General Unformation Theory: Some Macroscopic Dynamics of the Human Thinking Systems", *Information Processing & Management*, Vol.34, No.2/3 (1998), 275-290.

[21] Law, S. K., P. L. Nunez, A. F. Westdorp, A. V. Nelson, and K. L. Pilgreen, "Topographical mapping of brain electrical activity", *Proceedings of the IEEE Conference* (1991), 194-201.

[22] Montazemi, A. R. and D. W. Conrath, "The Use of Cognitive Mapping for Information Requirements Analysis", *MIS Quarterly*, Vol.10, No.1(1986), 44-55.

[23] Oard, D. W., "The State of the Art in the Text Filtering", *User Modeling and User-Adapted Interaction*, Vol.7, No.3(1997), 141-178.

[24] Oard, D. W. and G. Marchionini, "A conceptual framework for text filtering", Technical Report, CS-TR3643, 1996, University of Maryland.

[25] Schafer, J., J. Konstan, and J. Riedl, "E-commerce recommendation applications", *Data Mining and Knowledge Discovery*, Vol.5, No.1&2(2001), 115-153.

[26] Uckun, S., C. Ruokangas, P. Donohue, and S. Tuvi, "AWARE: Technologies for interpreting and presenting aviation weather information", *IEEE Aerospace Conference Proceedings*, Vol.2(1999), 443-449.

[27] Rodhain, F., "Tacit To Explicit: Transforming Knowledge Through Cognitive Mapping - An Experiment", *Proceedings of the ACM SIGCPR Conference on Computer Personnel Research*, (1999), 51-56.

[28] Ruch, P., R. Baud, G. Antonie, C. Lovis, A. Rassinoux, and A. Riviere, "Using Part-of-Speech and Word-Sense Disambiguation for Boosting String Edit Distance Spelling Correction", *Artificial Intelligence Medicine: 8th Conference on AI in Medicine in Europe*, (2001), 249-257.

Abstract

An Information Filtering System Using Cognitive Mapping

Jinhwa Kim* · Seunghun Lee* · Hyunsoo Byun*

Information filtering systems, which are designed for users' needs, do not satisfy user's diverse requests as their filtering accuracy is unstable sometimes. This study suggests an information filtering system based on cognitive brain mapping by simulating the processes of information in human brain. Compared to traditional filtering systems, which use specific words or pattern in their filtering systems, the method suggested in this article uses both key words and relationships among these words. The significance of this study is on simulating information storing processes in human brain by mapping both key words and their relationships among them together. To combine these two methods, this study finds balances in representing two methods by searching optimal weights of each of them.

Key words : Information Filtering, Filtering Agent, Contextual Filtering, Brain Mapping, Cognitive Mapping

* College of Business Administration, Sogang University