학술정보시스템의 온라인 인터페이스에 관한 연구 : 적응형 인터페이스를 위한 정보조직 및 활용

A Study on Online Interface for Research Information Systems : Information Organization for Adaptive Interface

김 미 현(Mi-Hyeon Kim)*

목 차

1. 서론
2. 적응형 인터페이스
2. 1 정보탐색과정
2. 2 정보필터링 시스템
2. 3 기계학습을 통한 적응성부여
2. 4 적합성피드백
2. 5 이용자 프로파일 구축방법
3. 지능형 정보검색 시스템의 사례
4. 이용자 지향적 적응형 시스템의 구축
4. 1 이용자 프로파일
4. 2 적합성 피드백과 이용자 피드백
5. 결론

초 록

이 연구는 이용자들이 문제해결을 위해 표현한 정보요구뿐 아니라 표현되지는 않았지만 내재된 정보요구를 모두 만족시키고, 이용자의 수준과 선호도에 따라서 시스템이 적응적으로 대처할 수 있도록 하는 적응형 인터페이스의 구성의 한 방안으로 기계적 학습과 결정트리를 이용한 이용자 프로파일 구축, 결합벡터를 이용한 적합성 피드백, 그리고 이용자 피드백을 활용한 지능적이고 이용자 지향적인 시스템을 제시하고자 한다.

ABSTRACT

This study is to contribute to develop adaptive information systems meeting inside information needs as well as represented information needs, and dealing with every levels of users and user's preferences. Also, this study is to present a method of developing adaptive information system through developing user profile using machine learning and decision tree, applying relevance feedback using merged vector, and applying user feedback.

* 첨단학술정보센터 (Korea Research Information Center) - Post Doc.
접수일자 1998년 5월 18일

1. 서론

고도의 정보화 사회에 있어서 국가의 경쟁력을 갖추는데 중요한 비중을 차지하는 것이 학술정보이다. 이렇게 정보에 대한 필요성이 증가함에 따라 국내 및 해외 정보를 신속히 수집, 보급하기 위해서 많은 정보검색 엔진들이 개발되어 왔다. 특히, 최근에는 컴퓨터 자체능력, 정보통신 기술, 통신망 등의 발달로 인해 학술정보의 효율적인 검색과 원거리 접근의 유용성을 증가시키기 위하여 전자도서관 구축이 일반화되고 있는 추세이다. 따라서, 시스템과 일반 이용자들간의 커뮤니케이션을 원활히 할 수 있는 전자도서관에서의 인터페이스 개발에 더욱 박차를 가하게 되었고, 이용자의 정보검색을 효율적으로 지원할 수 있는 이용자 지향적인 인터페이스 개발에 관한 관심 역시 증가하게 되었다.

정보검색에 있어 효율성을 고려한다는 것은 최신의 적합한 정보를 정보의 형태에 상관없이 원하는 시기에 제공한다는 것을 말하며, 적합한 정보란 문제의 해결을 위하여 발생된 이용자들의 요구를 만족시키고자하는 정보를 말한다. 이러한 정보는 이용자들이 문제해결을 위해 표현한 정보 요구를 만족시키고자 할 뿐 아니라 이용자들이 표현하지는 않았지만 내재해 있는 요구를 만족시키기 위한 정보를 모두 포함한다. 적합한 정보 검색에 관한 연구는 정보 검색 분야의 근간을 이루어왔으며, 이 중 이용자와 시스템간의 상호작용과 이용자 모델에 관한 연구가 효율적 정보검색을 위한 시스템 개발을 위해 수행되어 온 것이다. (Saracevic, Spink, and Wu, 1997)

현재는 인터넷의 일반화로 인해 다양한 유형의 이용자들이 수많은 정보원으로 접근할 수 있게 되었으며, 이로 인해 각종 정보검색을 위한 싸이트나 전자도서관들이 구축되고 있다. 이러한 많은 정보검색을 위한 싸이트들과 전자도서관들이 구축되면서 처음에는 기능적인 개발에 치중하였으나 시간이 흐르면서 이용자의 이용을 용이하게 하기 위한 이용자 지향적이고 지능적인 인터페이스 환경조성이 필요하게 됨에 따라 이에 많은 노력을 기울이고 있다. 특히 미래의 전자도서관은 많은 양의 정보 속에서 적합한 정보를 효과적으로 제공할 수 있도록 해야하는데, 이를 가능하게 하는 시스템에 대하여 "knowbot"라는 용어를 사용하였다 (Lesk, M., Fox, E., and McGill, M., 1993). 이는 단순히 자동화된 검색시스템만을 의미하는 것이 아니라, 이용자들을 이해하고 이용자들의 잠재적 요구와 표현된 요구사이의 차이를 파악하여 이용자들이 의식하지 못하더라도 컴퓨터에 지적인 처리과정을 부가시켜 이용자의 부담을 줄이는 시스템을 의미한다. 표현된 질의에서 잠재적 요구까지 파악하는 것은 처리과정에 있어 쉬운 일이 아니다. 이를 위해 정보처리과정과 함께 이용자와 직접 상호작용이 이루어지는 인터페이스 관련 영역에서 많은 연구들이 이루어지고 있지만, 아직까지는 더욱 많은 연구가 필요로 되고 있다. 적응형 사용자 인터페이스도 이러한 이용자 지향적인 지능형 정보검색시스템에 관한 연구분야중의 하나로 정보이용자의 수준과 선호도에 따라서 시스템이 적응적으로 대처할

수 있도록 하는 인터페이스를 말한다.

본 연구에서는 적응형 인터페이스의 구성을 위한 여러 처리과정들과 요소들을 살펴보고 이러한 과정 속에서 이용자들의 선호도가 반영되는 방법에 관하여 고려해 보고자 한다. 일반적으로 이용자들의 초기 검색에서 원하는 정보를 만족스럽게 얻기는 어렵다. 따라서 탐색을 수행해 나가면서 여러 단계의 수정작업들을 거치고 원하는 정보를 얻기 위해 조정작업을 하게 되는 것이다. 본 연구에서는 정보탐색과정 중에서 적응형 사용자 인터페이스를 통하여 이용자가 비효율적 탐색을 수행하였을 경우 좀더 효율적 검색을 할 수 있도록 지원하는 시스템 개발에 도움을 주고자 한다. 즉, 이용자가 적절한 정보검색을 수행하지 못했을 경우 이용자가 질의어를 재구성하는 과정을 거치지 않고 적합성 피드백과 이용자 피드백, 그리고 이용자 프로파일의 관리를 통하여 질의어 재구성에서 발생되는 반복적인 오류를 방지하고 이용자들의 표현되지 않은 실제적 요구에 맞는 효율적 정보검색을 할 수 있도록 하는 시스템의 구축에 도움을 주고자 한다. 따라서 본 연구는 사용자들의 관심분야와 선호도에 따라 이용자들이 인식을 못한다 하더라도 탐색을 효율적으로 구성할 수 있도록 도와주고 이용자들에게 적합한 정보를 제공하기 위한 적응형 인터페이스 구성에 대해 살펴보기로 한다.

2. 적응형 인터페이스

최근의 인터넷 환경에서의 검색엔진을 통한 서비스는 과다한 불필요한 정보들과 함께 제공하게 되므로 많은 이용자의 시간과 노력을 요구하게 되고 이용자를 더욱 혼란스럽게 할 소지가 있다. 이에 따라, 지능형, 적응형 에이전트들에 대한 연구가 활발히 이루어지고 있는 것이다. 적응형 시스템이라는 것은 다양한 분야에서 사용되는 용어이지만, 포괄적인 개념을 살펴보면 설계자가 예상치 못한 상황이나 문제에 대하여 자체반응을 하여 해결책을 제시할 수 있는 시스템을 말한다. (Prechelt, 1994) 또한 에이전트라는 용어는 특정목적을 수행하고자 하는 이용자를 대신하여 작업을 수행하며, 독자적으로 존재하지 않고 어떤 환경의 일부로서 운영되는 시스템이다. 또한 스스로 경험과 환경의 변화에 따라 학습하는 기능을 가지는 자율적 프로세서를 의미한다. (최중민, 1997)

분산환경에서의 에이전트 구조로는 크게 3가지 유형의 에이전트가 있는데, 인터페이스 에이전트, 테스크 에이전트, 정보에이전트가 있다. 인터페이스 에이전트는 이용자와의 상호작용을 하며 이용자의 질의를 받아 분석하고 결과를 보여주는 역할을 수행한다. 좀더 지능을 가진 인터페이스 에이전트는 이용자의 습성과 기호등의 정보를 수집하여 이를 바탕으로 시스템의 조정작업에 활용하는 기법을 가진다. 테스크 에이전트는 주어진 작업에 대한 영역지식과 함께 다른 테스크 에이전트나 정보에이전트의 능력 정보 등을 가지고 이용자가 요구한 작업을 실제로 수행하는 에이전트이다. 정보에이전트는 여러 곳에 흩어져 있는 이형질의 정보소스를 지능적으로 접근할 수 있는 기능을 제공한다. (최중

민, 1997) 이러한 세 가지 유형의 에이전트 중에서 본 연구는 이용자와 상호작용을 하여 이용자의 기호나 행태에 관한 정보를 수집하여 조정하는 이용자 지향적 인터페이스 에이전트에 관한 응용 방법에 대해 고려해보고자 한다.

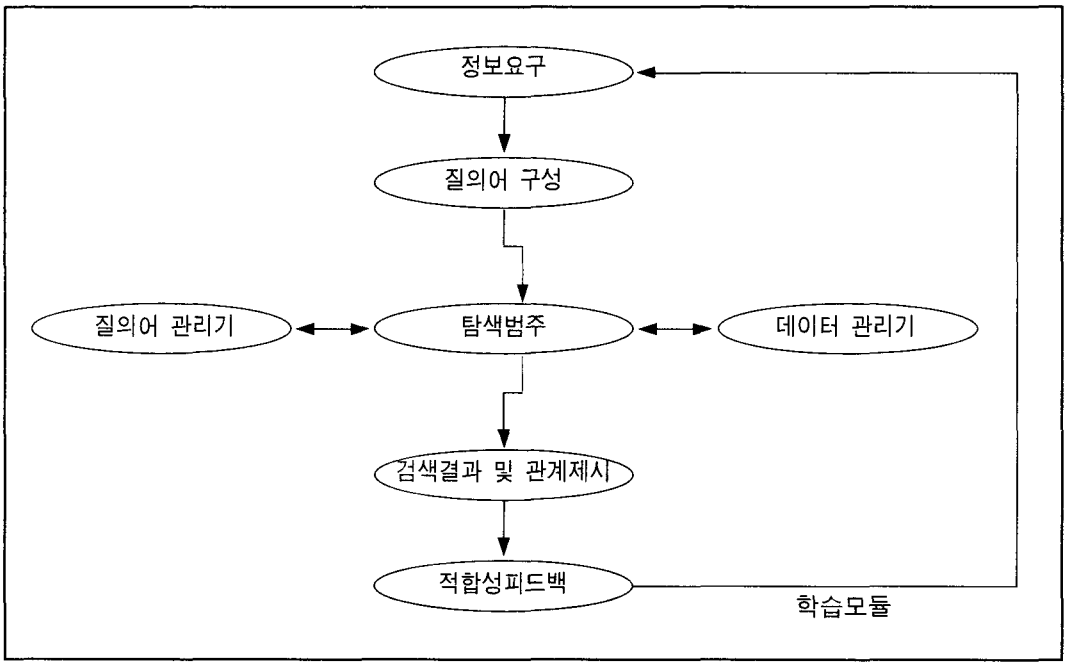
적응형 인터페이스를 구축한다는 것은 Russel과 Wefald (1991)가 언급하듯이 어떤 완벽한 시스템을 구축 한다기 보다는 이상적으로 상식적인 판단을 하는 에이전트로서 언제나 문제해결을 위한 올바른 접근을 하는 시스템 구축을 의미하는 것이다. 따라서, 지능형 인터페이스를 위해서는 설계의 원칙들과 기술적인 요소들이 조화를 이루고 이를 바탕으로 이용자의 관심분야나 탐색과정들을 이해하며, 필터링이나 기계적 학습을 통해 다음의 검색에서 보다 효과적이고 정확한 정보검색을 수행할 수 있는 시스템을 설계하도록 해야한다. 일반적으로 적응형 인터페이스로 구축되는 시스템의 기본구조를 살펴보면 〈그림 1〉과 같다.

- 질의어 관리기 : 자연어나 다른 모든 유형의 언어들을 번역하는 프로그램들을 포함하며, 탐색어에 대해 어떤 파일을 검색하고 어떤 조건에서 어떤 용 어들을 검색할지를 구분해놓은 탐색범주로 번역한다.

- 탐색범주기 : 질의어에 대하여 검색할 파일들과 조건들을 정의해놓은 시스템이다.

- 데이터관리기 : 탐색범주에서 정해놓은 범위에서 탐색과 검색을 수행하게 하는 시스템

- 적합성피드백 : 실제 상황에서 각기 다



〈그림 1〉 적응형 인터페이스

른 유형의 자료들 중 어떤 정보를 선호하는지에 관해 학습하며 이용자 피드백과 실제상황 범주의 적용에 따라 가치 값의 조정을 한다. 그리고, 상호작용의 과정을 통해 가장 높은 가치 값을 가진 업무에 가장 합당한 검색을 수행한다.

- 학습모듈 : 이용자 프로파일을 포함하여 등록된 이용자의 탐색과정을 모니터하고 반복되는 탐색행위의 패턴을 자동화한다. 예를 들어, 정보검색 필터링 에이전트는 사용자가 주로 읽는 논문들의 패턴을 학습하여 그와 비슷한 논문이 발견되면 사용자에게 제공할 수 있다. 또한 이용자의 실제 탐색요구를 파악하여 자동적으로 용어의 가중치를 선정한다.

일반적으로는 이와 같은 적응형 인터페이스의 구조를 가지는데, 이러한 구조를 이해하고 보다 검색의 효율성을 높이는 인터페이스의 구축을 위해 다음의 몇 가지를 자세히 살펴보고자 한다.

2. 1 정보탐색과정

이용자의 정보검색 행태를 이해하기 위하여, 우선적으로 정보탐색과정을 살펴보아야 한다. 이용자측면의 정보탐색과정의 근간을 이루는 중요한 학습이론들중 John Dewey, George Kelly, 그리고 Jerome Bruner에 의한 이론들을 정리하여 비교를 해보면 다음과 같다.

이러한 세 가지 이론들은 모두 다른 용어들을 사용하고 있으나 결론적으로는 개인적인 정보요구에 대한 문제해결을 위해 검색행

〈Table 1〉 정보요구 형성의 단계와 정의 (Kuhlthau, 1992)

Dewey - 사고의 단계 (Phrases of Reflective Thinking)

단계들	정의
Suggestion	Doubt due to incomplete situation
Intellectualization	Conceptualizing the problem
Guiding Idea (Hypothesis)	Tentative interpretation
Reasoning	Interpretation with more precise facts
Action	Idea tested by overt or imaginative action

Kelly - 정보요구 형성의 5단계 (Five Phases of Construction)

단계들	정의
Confusion and doubt	New experience
Mounting confusion and possible threat	Inconsistent/incompatible information
Tentative hypothesis	A direction to pursue
Testing and assessing	Assessing outcome of undertaking
Reconstructing	Assimilating new construct

Bruner - 정보요구 정립을 위한 과정 (The Interpretive Tasks)

단계들	정의
Perception	Encountering new information
Selection	Recognizing patterns
Inference	Joining clusters and categories
Prediction	Going beyond the information given
Action	Creating products of the mind

위를 취할 때까지의 과정을 처음 정보에 대한 요구가 발생한 후 실제로 검색활동을 시작할 때까지 5가지 단계의 사고과정을 거친다고 분석한 것이다. 그리고 이러한 사고과정을 거쳐 실제의 검색활동이 시작된 이후에도 다시 다음의 6단계의 과정을 거치게 된다. 그 첫 번째 단계가 연구활동의 시작 (task initiation)의 단계로서 탐색을 위한 정보의 요구가 시작되는 단계이고, 두 번째 단계는 주제선정 (topic selection)의 단계로 발생된 정보요구가 아직 정립화 되어 있지 못하고 일반적인 주제를 선정하는 단계이다. 세 번째 단계는 사전탐색 (prefocused exploration)의 단계로서 일반적 단계에서 좀더 세밀하게 주제 파악을 하고 좀더 확실하게 정보요구를 성립하는 단계이다. 네 번째 단계는 정보요구의 확립 (focused formulation)의 단계로서 탐색과정에 있어 정보요구의 불확실성이 감소되고 정보요구를 확실히 파악하여 주제어를 선정하는 단계이다. 다섯 번째 단계는 정보수집의 단계로서 선정된 주제어를 가지고 정보수집이 효과적이고 효율적으로 이루어지게 하는 단계이다. 그리고 마지막 단계는 탐색종료의 단계로서 탐색이 성공적이든 불만족스럽든 또한 어떠한 이유로 탐색을 중지하든 탐색을 마치는 단계이다.

정보서비스를 효과적으로 수행하기 위해서는 이러한 정보탐색과정의 단계들을 이해하고 다음의 세 가지 요소들을 고려해야 한다 :

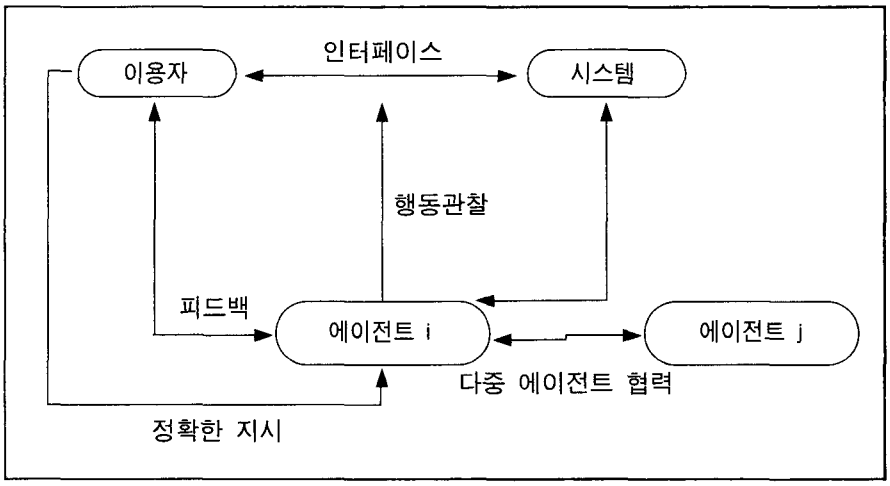
1. 탐색과정의 가시화 (charting) : 이용자의 전체 탐색과정을 가시화하고 각 단계에서의 탐색행동을 알고 있어야한다.

2. 질의어 구성의 지원 (composing) : 이용자에게 탐색주제에 있어 명확히 하도록 도와주어야 한다.

3. 내재적 요구파악 (converting) : 이용자들이 실제요구와 표현된 요구의 차이를 이해하고 탐색과정에 있어서의 불확실성을 감소시켜야 한다. (Kuhlthau)

이용자에 대한 정보를 수집하는데는 다음의 네 가지 정보원들이 있다. 반복적인 이용자 행동양식의 관찰, 이용자의 피드백, 이용자의 직접적 상호작용을 통한 학습, 그리고 다른 에이전트와의 협조가 그 방법들이다. (Akoulchina, and Ganascia, 1997)

이렇게 학습이론과 정보탐색과정에서 볼 수 있듯이 이용자의 정보요구를 탐색 시 이끌어내기는 쉽지 않은 문제이다. 따라서, 이용자의 정보에 대한 불확실성을 감소시키기 위해 이용자를 이해하는 지능적인 정보검색 시스템의 개발이 요구된다고 할 수 있다.



〈그림 2〉 학습 인터페이스 에이전트가 지식을 습득하는 4가지 정보원

2. 2 정보필터링 시스템

정보검색시스템을 구축하는데는 수많은 이론들과 시도들이 있어왔는데, 그중 이용자의 선호도를 반영하면서 이용자의 지적부담을 감소시키고, 이용자로 하여금 방대한 양의 정보 속에서 적합한 정보만을 검색할 수 있도록 하는 지능형 정보검색시스템을 구축하기 위해 많은 노력을 기울여 왔다. 그중 필터링(Filtering) 시스템은 이용자들의 요구나 선호도를 저장하고 있는 프로파일을 바탕으로 필터링을 함으로서 검색된 결과의 수를 줄이는 방법이다. 필터링은 다양한 방법을 통해 이루어질 수 있지만, 그중 기계적 학습을 통해 이용자의 프로파일을 갱신하고 다시 필터링을 하여 이용자에게 보다 적합한 정보를 제공하는 것이 가장 일반적인 방법이다.

그 예로, Automated collaborative filter-ing 시스템은 인간의 지식과 기계의 스피드를 이용하여 컴퓨터가 이용자들의 검색행위를 통해 비슷한 관심분야를 가지는 이용자들을 분석하여 이들의 정보유통을 원활히 하고 좀 더 적합한 정보제공을 하고자 한 것이다. 또한, Group Asynchronous Browsing 시스템 역시 인간의 지식을 활용하는 시스템으로 이용자들의 북마크나 핫 리스트 파일들을 수집하고 통합하는 서버를 개발하여 비슷한 연구영역을 가진 이용자들의 정보원에 대한 지식을 활용하고자 하였다. (Meadow, 1997)

2. 3 기계학습을 통한 적응성 부여

기계학습에 기반을 둔 방법은 최소한의 배경지식으로 사용자의 행위를 학습함으로써 사용자를 돕는데 필요한 지식을 얻는 방법으로서 일반적으로 4가지의 방법을 통해 학습한다. 그 첫째 방법은 사용자의 행위를 관찰함으로서 반복되는 행위의 패턴을 자동화하는 것이다. 예를 들어, 뉴스필터링 에이전트는 사용자가 주로 읽는 기사의 패턴을

학습하여 그와 비슷한 기사가 발견되면 사용자에게 그 기사를 제공하는 것이다. 둘째 방법은 사용자의 피드백을 통한 학습으로 피드백에는 간접적 또는 직접적 피드백이 있다. 간접적 피드백은 에이전트가 제의하는 행위를 사용자가 무시하고 다른 행위를 취할 때 이를 저장하는 것이며, 직접적 피드백은 사용자가 "이런 행위를 하지마"와 같이 직접적 피드백을 줌으로써 에이전트의 행위를 바꾸는 것이다. 셋째 방법은 사용자가 의도적으로 사례를 제시함으로써 훈련을 통하여 에이전트를 학습시키는 것이다. 사용자는 에이전트에게 가상의 사건, 상황에 대한 사례를 제시하고 그러한 상황에서는 무엇을 해야할 것인지를 보이면서 에이전트를 훈련시키는 것이다. 예를 들어, 특정한 사람에게서 온 메일을 특정한 폴더에 저장하는 예를 보임으로써 에이전트를 훈련시킬 수 있다. 마지막 방법은 다른 에이전트의 충고를 통하여 필요한 지식을 습득해 나가는 방법이다. (박혜경, 1997) 에이전트에서 해결할 수 없는 문제에 있어서는 다른 에이전트들을 이용하여 그 해결방법과 과정을 학습하는 것이다. 이러한 기계학습 시스템구축방법에는 사례기반 학습시스템과 규칙기반 학습시스템으로 나누어 볼 수 있다.

2. 3. 1 사례기반 학습시스템

사례기반 학습방법은 특정 개인 사용자로부터 관심 키워드를 받거나 하나의 탐색결과를 하나의 실례로 저장, 평가하여 학습하는 방법이다. 사례기반 학습방법은 여러 분야에 관련이 있는 연구영역이나 관심 분야가 한정적이지 않다면 다소 어려움이 따를 수 있으나, 이용자의 관심분야가 비교적 명확한 경우에는 효과적으로 학습할 수 있는 방법이다. (이말례와 남태우, 1997)

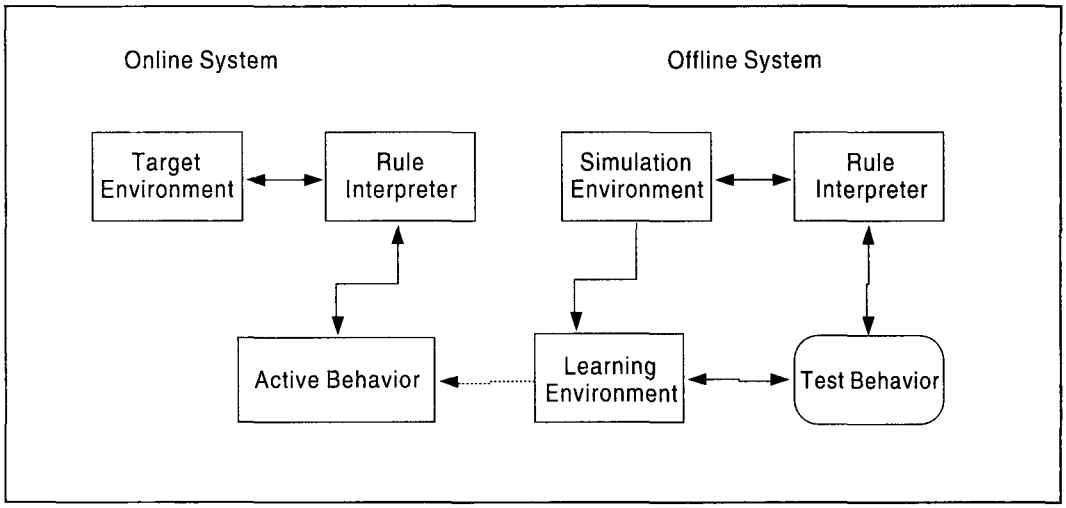
2. 3. 2 규칙기반 학습시스템

이는 번역기가 현재의 상황이나 해결해야 할 문제를 인식하고 기존에 관찰된 행동양식에 기반을 두고 문제해결에 합당하다고 여겨지는 규칙을 찾아 이를 행동으로 옮기는 것이다. 이러한 시스템의 구조를 보면 〈그림 3〉과 같다. (Schultz, 1994)

2. 4 적합성 피드백

적합성 피드백은 기계적 학습과 필터링 시스템을 이용하여 시스템을 구축하는데 주로 이용되는 방법이다. 피드백이라는 과정을 통한 결과물은 미래의 탐색과정을 통제하기 위해 정보를 표현하는 것이다. 피드백으로 가능한 정보들은 파일이나 데이타베이스 선정, 용어탐색이나 브라우징, 레코드 탐색이나 질의어 구성, 레코드 디스플레이와 브라우징, 레코드 수집, 정보검색에 대한 정보요청, 커뮤니케이션 범주설정, 그리고 탐색과정에 있어서의 패턴들이 있으며 보다 나은 검색을 위해 용어선정의 향상, 질의어구성의로직의 향상, 마지막 검색 결과의 향상, 정확성, 검색률, 그리고 전체적 효용성의 향상을 통해 조정될 수 있다.

적합성피드백에서 일반적으로 많이 연구되는 분야는 다음의 두 가지가 있다. 첫째는 질의어에 따른 적합 문헌과 비적합 문헌들에



〈그림 3〉 규칙기반 학습방법 구조 (Schultz, 1994)

서의 용어의 분포에 따라 질의어의 가중치를 다시 주는 방법이고, 두 번째로는 실제 질의어의 조정과 질의어에서 사용된 용어의 피드백에 관련된 방법이다. (Frakes, 1992)

2. 5 이용자 프로파일 구축방법

이용자 프로파일은 이용자의 선호도를 반영하기 위한 것으로 많은 시스템들이 기계적 학습방법과 정보필터링 방법을 조합하여 양질의 이용자 프로파일을 구축하기 위해 시도하고 있는데, 그중 많이 이용되고 있는 방법들로는 tf-idf(term frequency-inverse document frequency weighting) (Salton and McGill, 1983)와 Bayesian Classifier (Duda & Hart, 1973), 결정트리 (Decision Tree) 들이 있다.

2. 5. 1 td-idf (Term Frequency-Inverse Document Frequency weighting)

이는 특정 문헌에 자주 나오지만 다른 문헌에는 잘 나오지 않는 용어가 그 문헌의 topic에 가장 적합한 것이라는 개념을 가진 방법이다. (수식1)에 의해 계산된 용어의 빈도는 특정문헌 내에서 나타나는 확률을 나타내는 중요도이다.

$$Weight_{ik} = Freq_{ik}/NoWords_{si} \quad \text{(수식 1)}$$

$Freq_{ik}$: 특정문헌 i에 용어 k가 나타나는 빈도수
$NoWords_{si}$: 문헌 i에 있는 용어의 수

td-idf 에서는 장서의 문헌들간에 주제 구분을 해줄 수 있는 용어에 가중치를 준다.

특정 용어 k가 다른 문헌에 비해 특정문헌에 얼마나 나타나고 있는지에 대한 가중치를 계산하는 수식은(수식 2)와 같다.

$$Weight_{ik} = Freq_{ik} * [\log_2 n - \log_2 DocFreq_k + 1]$$

(수식 2)

n : 총 문헌의 수

$DocFreq_k$: 용어 k가 나타나는 문헌의 수

또한 문헌의 주제 분류들 중 선호도를 나타내는 용어의 적합성 판정은 (수식 3)에서와 같이 계산된다. 이는 주제분류 c에 속해있는 문헌 중 용어 k를 포함하는 문헌들과 주제분류 c에 속하지 않은 문헌들 중 용어 k를 포함하는 문헌들을 비교함으로써 주제분류 c에 대한 용어 k 에 대한 용어의 중요도를 나타내고자 한 것이다. (Salton and McGill, 1983)

$$TermRelevance_{kc} = [r_{kc}/(R_c - r_{kc})] / [s_{kc}/(I_c - s_{kc})]$$

(수식 3)

r_{kc} : 용어 k를 포함하는 class c에 속하는 문헌의 수

R_c : class c에 있는 문헌들의 수

s_{kc} : 용어 k를 포함하는 class c에 속하지 않는 문헌의 수

I_c : class c에 포함되지 않는 문헌의 수

따라서 td-idf는 사용자의 관심도를 나타내는 벡터와 이용자들이 선택한 용어의 벡터와 이용자들이 선택하지 않은 용어의 벡터사이의 유사도를 파악하여 이용자의 관심에 적합한 용어에 대한 값을 결정한다. 이렇게 문헌들에 대한 학습을 함으로서 지식을 얻어, 이용자에게 현재 검색하는 분야에서 가장 적합한 문헌들을 제시하는 것이다.

2. 5. 2 Bayesian Classifier

Bayesian Classifier는 분류를 위한 조건확률을 이용한 방법이다. 이는 예 j가 주어진 속성값 (attribute value)에 따라 주제분류 C i에 속할 수 있는 확률을 결정하게 하는데 이용된다. 예제 j가 속성값 A_1을 가지고 주제분류 C_i 에 속하는 확률은 (수식4)와 같다.

$$P(C_i|A_1 = V_{1j} \&...\& A_n = V_{nj})$$

(수식4)

그러나 만일 속성 값이 독립적이라면 이 확률은 수식 5와 같다.

$$P(C_i)\prod_k PROD\ P(A_k = V_{kj} \mid C_i)$$

(수식 5)

$P(A_k = V_{kj} \mid C_i)$ 와 $P(C_i)$는 학습 데이터에서 습득될 수 있다. 가장 합당한 분류에 할당하기 위해서 각 분류의 확률이 계산되고 가장 높은 확률을 가진 분류에 할당된다. (Pazzani, Murumatsu, and Billsus, 1996)

2. 5. 3. 결정트리 (Decision Tree)

결정트리 방법은 학습에 있어 문헌의 중요한 개념들을 경험적인 방법으로 구성하여 문헌을 그 개념에 따라 범주화하도록 함으로서 문헌을 그룹화 하는 것이다. 따라서, 결정트리 방법은 개념의 그룹화가 목적이라고 할 수 있고, 이를 위해서는 인간의 경험이나 선호도에 따라 개별화하여 그룹 지을 수도 있고 수학적 방법을 도입하여 적합성 여부를 고려하여 그룹 지을 수도 있다.

결정트리는 객체를 분류하는 규칙들이 트리의 형태로 나타나며 각 단말노드는 하나의

범주에 할당된다. 테스트의 결과에 따라 또 다른 서브결정트리가 형성되는 것이다. 모든 객체들은 각각 하나의 범주에 속하게 되며 객체들은 결정트리의 루트에서 단말노드까지의 경로를 따라서 분류되어진다. 객체는 특성 값(attribute value)의 집합으로 나타나며 이 특성 값은 이산치(discrete value)내지는 연속된 값이다. 객체들의 집합 S로부터 각각의 범주로 분류하기 위한 결정트리를 생성하는 처리단계는 다음의 두 단계로 이루어진다.

단계 1 : S에 있는 모든 객체들이 같은 범주 Ci에 속하면 Ci의 이름을 가진 하나의 단말노드로 구성된 결정트리가 생성된다.

단계 2 : 그렇지 않으면, 테스트 T의 결과 O1, O2, ... On에 따라 S는 부분집합 S1, S2, ... Sn으로 나뉘어진다. Si에 있는 모든 객체들은 Oi의 결과를 가진다. T가 결정트리의 루트가 되고 단계 1로 가서 서브 결정트리의 Si에 대해서 반복적으로 수행한다.

그러나, 결정트리를 반복함으로써 발생되는 지나친 확장을 막기 위하여 분류정지기준(stopping criterion)을 적용하여 단말노드에 적정수준 이하의 객체가 속한 경우 더 이상의 분류를 행하지 않는다. (양수연, 1994) Bloedorn, Mani, 그리고 MacMillan (1997)은 결정트리를 적용한 시스템과 td-idf를 적용한 시스템을 비교 실험한 결과 결정트리를 적용한 방법이 더 정확하고 포괄적인 학습프로파일을 제공함을 보여주었다.

3. 지능형 정보검색시스템의 사례

3. 1 InfoFinder

InformationFinder는 이용자에 의한 분류나 이용자들로부터의 일련의 메시지들을 학습하는 지능형 에이전트이다. InformaionFinder는 수학적 계산보다는 경험적인 방법을 이용하여 많지 않은 문헌들을 기반으로 일반적 탐색범위를 학습할 수 있도록 하였으며, 각각의 이용자 범위에 대한 결정트리를 학습할 수 있도록 하였다. 이러한 결정트리는 쉽게 탐색질의 구문으로 전환된다. 이용자는 검색된 문헌에 대해 만족하는지 불만족 하는지를 나타낼 수 있도록 브라우저에서 웃는 얼굴이나 찡그린 얼굴의 아이콘을 선택하게 되고, 이용자로 하여금 그 주제에 대한 분류를 하게 한다. 이러한 분류는 각각 이용자에 대한 정보로 인식되어 이용자와의 상호작용에 이용된다. 이용자가 여러 문헌들의 특정분야에서 적합성을 선택한 뒤 InfoFinder는 이러한 분류를 질의어 구문을 학습하는데 이용된다. 그리고 이러한 질의어 구문은 그의 관심분야에 부응하는 자에 대한 새로운 메시지를 보내는데 이용된다. 에이전트는 이를 다음의 세 가지 단계로 실행한다.

1. 각 문헌에서 중요한 구절을 추출한다.
2. 추출된 구절이 속한 각 분야에 대한 결정트리를 학습한다.
3. 각각의 결정트리는 일반적인 탐색엔진을 위해 불리안 탐색질의어 구문으로 전환 된다. (Krulwich and Burkey, 1996)

3. 2 DR-LINK (Document Retrieval Using LINguistic Knowledge)

이러한 기계적 학습을 통한 지능형 인터페이스를 가진 시스템 중 DR-LINK (Document Retrieval Using LINguistic Knowledge) 시스템은 표현된 query의 실제 요구파악을 위해 자연어 처리를 통하여 인간의 개입 없이 지능형 검색을 가능하게 하고자 하여 ARPA (Advanced Research Projects Agency)에서 개발한 시스템으로 통계적 확률과 법칙들을 학습하여 검색률과 정확률이 가장 높은 검색결과를 제공하고자 한 시스템이다.

DR-LINK는 모든 질의 표현의 수준을 어형론상으로, 어휘적으로, 구문론적으로, 어의론 적으로, 그리고 구어론적으로 모두 분석하여 문헌들을 검색하는데 이는 다음의 4가지 방법을 이용하여 수행된다. 우선 이용자의 정보요구를 질의어와 문헌에서 사용되는 용어들의 의미파악을 통해 정확히 이해하고자 하며 이용자의 다양한 적합성의 범위를 파악한다. 그리고 다양한 정보원에서 표현될 수 있는 적합한 정보의 복잡성을 이해하도록 한다. 또한 질의에 부응하는 문헌들을 검색하는데 단순히 질의어분석 뿐만이 아니라 개념적 수준에서 질의어를 분석하여 적합한 문헌들을 제공하고자 하였다.

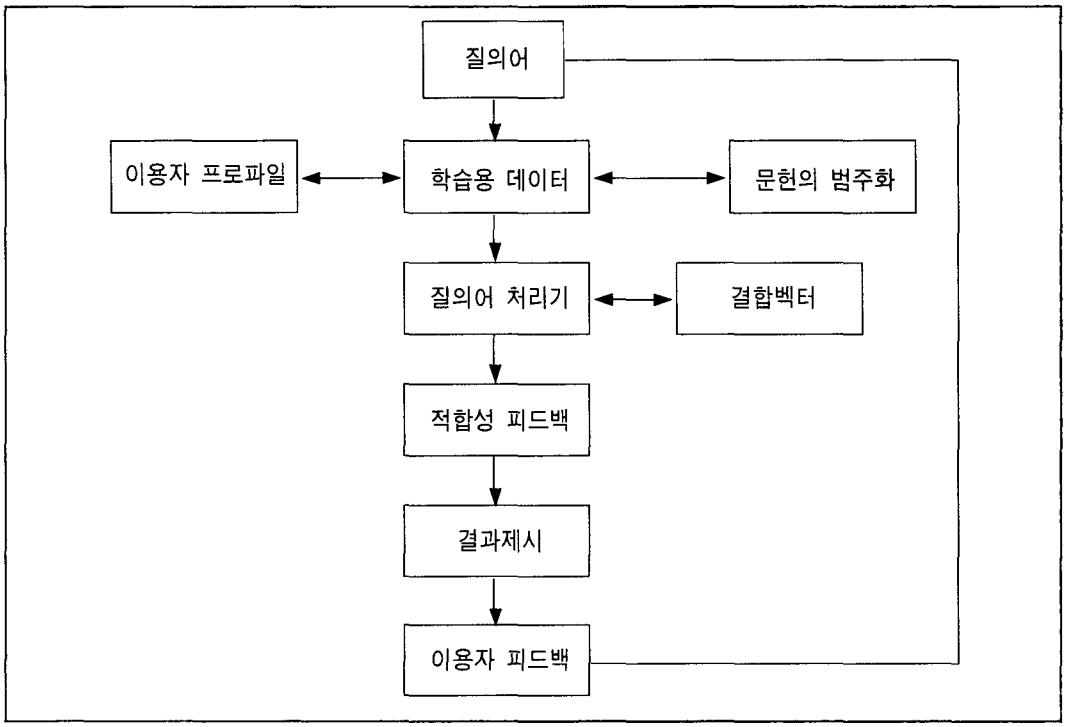
4. 이용자 지향적 적응형 시스템의 구축

지능적인 적응형 정보검색시스템의 구축을 위해서는 이용자에 관한 상세한 지식을 바탕으로 시스템 구축과 운영을 하는 접근이 가장 효과적이다. 앞에서도 논의되었듯이, 이용자의 정보요구가 확실하게 정립되기까지는 여러 가지 단계들을 거치게 된다. 이러한 이용자의 불확실한 요구까지 파악하며 효율적 정보탐색 과정을 수행하도록 하기 위해서는 이용자의 이용 행태에 관한 데이터 수집이 이루어져야 한다. 또한 적합성 피드백을 통하여 이용자의 탐색질의를 정확하게 구성할 수 있도록 하고, 이용자의 피드백을 이용하여 실제적 이용자의 선호도를 반영할 수 있어야 한다. 따라서 본 연구는 이용자 지향적인 적응형 시스템을 구축하기 위해서 다음과 같은 결정트리와 기계적 학습을 통한 이용자 프로파일 구축과 피드백 방법을 이용해 보고자 한다. (그림 4 참조)

4. 1 이용자 프로파일

이용자의 탐색 행태나 질의어 유형 파악 등을 통한 효율적인 이용자 관련 정보를 습득하기 위해서는 우선적으로 자세한 이용자 프로파일이 구축되어야한다. 이를 위하여 초기의 상세한 이용자 정보의 습득은 필수적이며 프로파일의 유지 또한 이루어져야 하는데, 본 논문에서는 기계적인 학습과 결정트리를 이용한 이용자 관심분야의 범주화를 수행함으로써 효율적인 프로파일 구성하는 방안에 관해 고려해 보고자 한다.

이용자의 등록 시 관심분야의 키워드를 입력하도록 하여 초기의 이용자 프로파일을



〈그림 4〉 이용자 프로파일 구성 및 이용자 선호도 적용

구축한다. 각각의 이용자는 등록 후 이용자 파일과 함께 인지되어 각 이용자들의 탐색 행태가 학습용 데이터베이스로 입력된다. 그리고, 임계값(threshhold)에 의한 확률 계산을 통해 결정트리로 범주화할 수 있는지를 학습하고 결정트리로 범주화 할 수 있는 정보는 이용자 프로파일의 갱신을 통해 프로파일에 포함이 되고 탐색어가 재구성된다. 그리고 임계값에 미치지 못하는 경우는 학습용 데이터베이스에서 추가 학습을 위한 정보로 남아 있게 된다. (그림 4 참조)

결정트리를 이용한 이용자의 관심분야 학습절차는 다음과 같다.

1. 학습용 데이터베이스에는 이용자의 탐색 히스토리와 관심분야, 이용자가 선호하는 용어들이 포함되어있으며, 이러한 데이터들의 학습을 통하여 개념적 결정트리 로 범주화하는 규칙을 얻는다.

2. 추가로 들어온 데이터에 대해서 범주정보와 확률 값을 얻는다.

3. 얻어진 확률 값이 임계값(threshhold)에 미치지 못할 경우 추가학습을 결정한다.

4. 추가학습을 통해 얻어진 확률 값이 임계값을 넘는 경우 이용자의 프로파일을 갱신한다.

4. 2 적합성 피드백과 이용자 피드백

이용자 피드백에서는 질의와 제시된 문헌들간의 적합성 분석에 실제적 이용자의 선호도를 적용하는 과정으로 이용자의 프로파일과 제시된 문헌들간의 관계를 학습함으로써 유형화를 조정하는 단계가 이루어진다. 본 연구는 실제 질의어를 재조정하는 적합성 피드백과정을 거쳐 보다 적절한 정보를 제공하고, 그 결과에 이용자의 선호도를 적용하고자 한다.

우선 적합성 피드백을 위하여 이기호등(1997)에 의해 테스트되어 가장 높은 검색효과를 보여주었던 결합모델인 Rocchio와 Pr__cl의 결합벡터를 이용하여 새로운 질의어 벡터를 이용하고 이에따라 검색결과를 제시한다. Rocchio는 1971년에 초기 질의 벡터와 적합 및 부적합 문서 벡터들의 벡터합에 의해 질의 벡터를 생산하는 것으로 다음과 같은 공식에 의해 계산된다.

$$Q_{new} = Q_{old} + \beta \cdot \sum_{i=1}^{n1} \frac{R_i}{n1} - \gamma \cdot \sum_{i=1}^{n2} \frac{S_i}{n2}$$

(수식 6)

Qold : 초기 질의에 대한 벡터
Ri : 적합 문서 i에 대한 벡터
Si : 비적합 문서 i에 대한 벡터
n1 : 적합 문서수
n2 : 비적합 문서수

Pr__cl 공식은 Croft & Harper (1979)에 의해 전형적인 확률검색 모델에 근거하여 개발된 것으로 색인어 ti의 적합성에 대한 가중치를 구하는 것이며 다음과 같은 식에 의해 가중치를 구한다.

$$W_{qi}' = \log \frac{P_i(1-q_i)}{q_i(1-P_i)}$$

$$P_i = \frac{\gamma_i + 0.5}{R+1}$$

$$q_i = \frac{n_i - \gamma_i + 0.5}{N-R+1}$$

(수식7)

γ_i : 색인어 ti를 갖는 적합 문서수
ni : 컬렉션내 색인어 ti를 갖는 문서수
R : 적합 문헌 총수
N : 컬렉션내 문서수

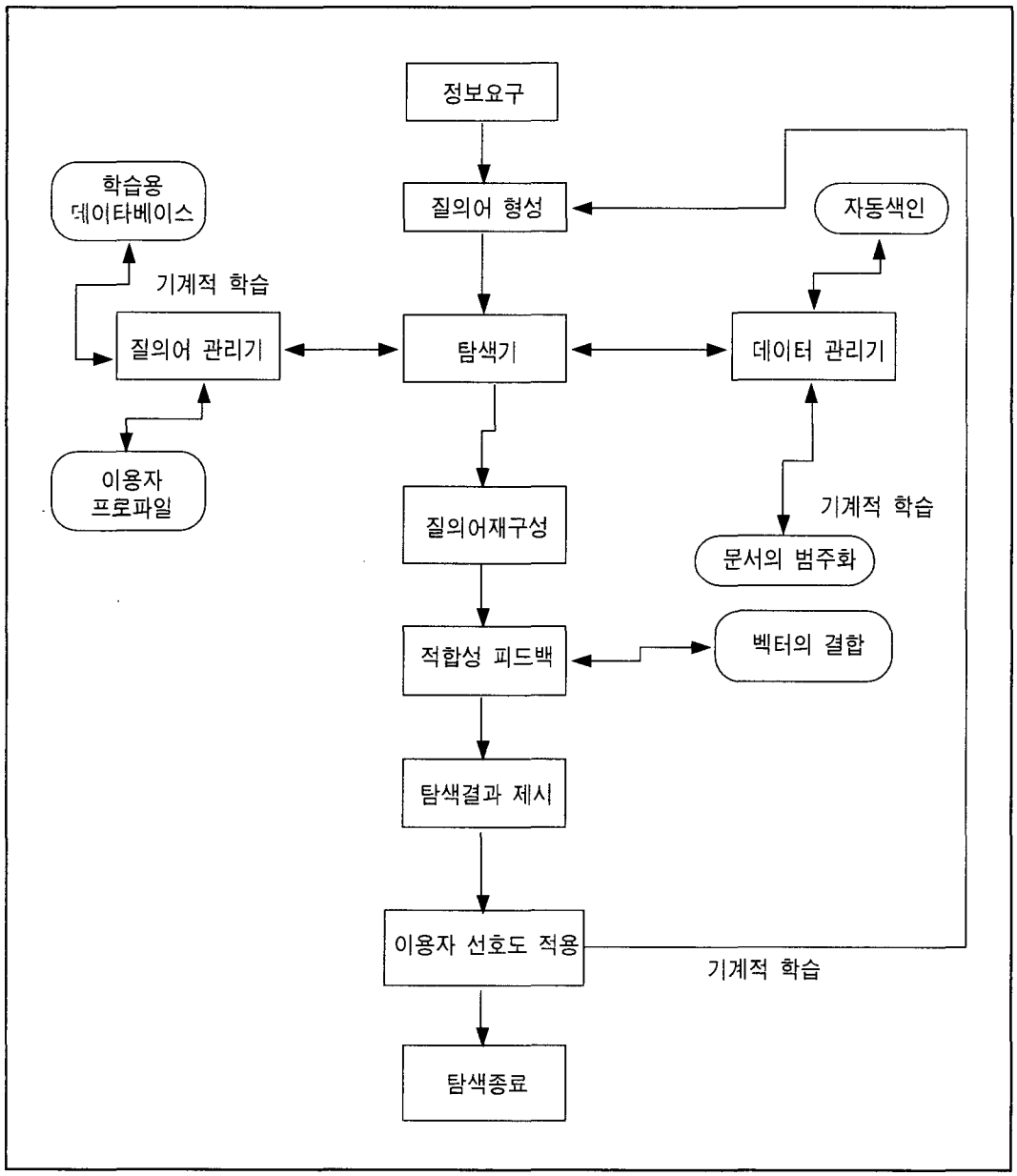
이와 같은 2개의 질의벡터를 결합하여 생성된 새로운 질의어 벡터에 의해 검색된 문헌들이 이용자에게 제공되고, 이용자들은 그 결과에 만족스럽지 못할 경우 보다 적합한 정보를 탐색하기 위하여 새로운 질의어를 입력하기보다 이용자 피드백을 하도록 권고한다. 만약 이용자가 피드백을 하고자하여 피드백을 시작하면 우선, 이용자는 검색된 문헌들 중 적합한 문헌을 표시하게 되고 시스템은 표시된 정보와 기존의 정보를 통합한다. 통합된 정보는 질의어의 범주를 재조정하는데 이용이 되고, 갱신된 범주로 문헌이 재검색 되는 것이다.

5. 결론

일반적인 적응형 인터페이스의 구조에 이용자의 선호도에 더욱 중점을 둔 이용자 지향적 적응형 인터페이스의 구조도는 다음과 같이 구성이 된다.

우선 정보의 요구가 일어나게 되면 먼저 시스템에 로그인 할 때 이용자의 기본적 프로파일 구성을 위한 형식을 작성하게 된다. 그리고 정보탐색을 위한 질의어가 형성될 때 질의어를 구성하고 있는 탐색어들과 탐색어들간의 구조는 분석이 되고 학습용 데이타베이스에 저장이 된다. 학습용 데이터베이스의 데이터들이 유형별로 범주화할 수 있는 임계값(threshhold)을 넘게되면 이용자 프로파일의 범주화에 대한 정보가 갱신되는 것이다. 이용자의 질의어는 결정트리의 구성 법칙에 따라 범주화가 이루어지며 이는 문서의 범주화와 비교되어 질의어가 조정되고 이는 결합벡터를 이용한 적합성 피드백 과정을 거쳐 문헌을 검색하게 된다. 검색된 결과는 이용자의 만족도에 따라 검색결과의 적합성이 판정되게 되고, 불만족스러울 경우 질의어의 재구성으로 인한 반복적인 오류를 방지하기 위하여 검색된 결과에 대하여 이용자가 피드백을 하도록 유도한다. 그리고, 이용자 피드백에 따른 정보에 따라 이용자의 프로파일이 다시금 조정되어 이용자가 원하는 정보검색을 위한 보다 정확한 질의어가 생성되는 것이다. 이러한 과정을 거치면서 이용자는 비효율적 탐색의 가능성을 내포한 질의어의 재구성 과정을 거치지 않고 이용자의 표현되지 않은 실제적 요구에 맞는 효율적 정보검색을 가능하게 하여 현재 많이 거론되고 있는 "infobot"에 대한 필요성에 부응하는 지능적인 정보검색시스템이 구축되는 것이다.

(그림 5 참조)



〈그림 5〉 이용자의 선호도를 적용한 적응형 인터페이스 구조도

참 고 문 헌

박혜경. 1997. "적응형 에이전트" 정보과학회지. v.15, n.3, pp.29-37.

양수연. 1994. 기계학습과 스크립트-기반 자연어 처리를 통한 정보추출. 포항 : 포 항공과대학 대학원.

이기호, 이준호, 이규철. 1997. "다중 질의 결합을 통한 검색 효과의 개선." 문헌정보학회지. v. 31, n. 3, pp.135-146.

이말례, 남태우. 1997. "학습방법을 이용한 지능형 웹 에이전트 시스템 설계" 정보관리학회지. v. 14, n.2, pp.285-301.

최중민. 1997. "에이전트의 개요와 연구방향" 정보과학회지. v.15, n.3, pp. 8-10

Akoulchina, Irina, and Jean-Gabriel Ganascia. 1997. "SATELIT-Agent : An Adaptive Interface Based on Learning Interface Agent Technology." Proceedings of the Sixth International Conference, UM97. New York : Springa Wien

Bloedorn, Eric, Inderjeet Mani, and Richard MacMillan. 1997. "Representational Issues in Machine Learning of User Profiles." AAAI Spring Symposium, 1997.

Bruner, J. 1975. Toward a Theory of Instruction. Cambridge MA : Harvard University Press.

Croft, W. B. and D.J. Harper. 1979. "Using Probabilistic Models of Document Retrieval without Relevance." Journal of Documentation. v.35, pp.285-95.

Dewey, J. 1933. How We Think. Lexington, MA : Heath & Company.

Duda, R. & Hart, P. 1973. Pattern Classification and Scene Analysis. New York : John Wiley & Sons.

Frakes, W. 1992. Information Retrieval : Data Structures and Algorithms. Englewood Cliffs, N.J. : Prentice Hall,

Kelly, G. A. 1963. A Theory of Personality : The Psychology of Personal Constructs. New York : W. W. Norton & Co.

Krulwich and Burkey. 1996. "Learning User Information Interests through Heuristic Phrase Extration." IEEE Expert/Intelligent Systems & Their Applications. v.12, n.5, pp.15-17.

Kuhlthau, Carol Collier. 1996. Seeking Meaning : A Process Approach to Library and Information Services. New Jersey : Ablex Publishing Co.

Meadow, Charles T. 1992. Text Information Retrieval Systems. San Diego : Academic Press, Inc.

Pazzani, M., J. Murumatsu, and D. Billsus. 1996. "Syskill & Webert : Identifying Interesting Websites." Proceedings of the National Con-

ference on Artificial Intelligence. Portland

Prechelt, Lutz. 1994. "A Motivating Example Problem for Teaching Adaptive System Design." ACM SIGCSE Bulletin v.26, n.4, p.25-28.

Rocchio, J. 1971. "Relevance Feedback in Information Retrieval." SMART Retrieval Systems : Experiments in Automatic Document Processing. Englewood Cliffs, N.J. : Prentice Hall Inc.

Lesk, M., Fox, E., and McGill, M. 1993. "A National Electronic Science, Engineering and Technology Library." Source Book on Digital Library. Fox, E. A., ed. Washington D.C. : National Science Foundation. pp.4-24.

Russel, Stuart and Eric Wefald. 1991. Do the Right thing : Studies in Limited Rationality. Cambradge : Mass : MIT Press.

Salton, G., and McGil, M. 1983. Introduction to Modern Information Retrieval. New York : McGraw-Hill.

Saracevic, Tefko, Amanda Spink, and Mei-Mei Wu. 1997. "Users and Intermediaries in Information : What are They Talking About?" Proceedings of the Sixth International Conference, UM97. New York : Springa Wien New York.

Schultz, Alan C. 1994. "Learning Robot Behaviors Using Genetic Algorithms"Procs. of the Inter-national Symposium on Robotics and Manufacturing, August 1994. p.14-8.