가중치 자동 조절을 이용한 매칭 에이전트

김동조* 박영택**

Matching Agent using Automatic Weight-Control

Dong-Jo Kim*, Young-Tack Park**

요 약

다차원의 속성들을 포함한 대용량의 데이터베이스 또는 정보 저장소의 데이터로부터 지식을 추출하고 이를 활용하기 위해서는 데이터 마이닝의 인공지능 기법 중 기계학습을 활용할 수 있다. 본 논문은 질의어를 바탕으로 각 속성들에 가중치를 적용하여 사용자가 원하는 데이터 집합을 분류하고, 사용자 피드백을 통하여 속성 가중치를 동적으로 변화시킴으로써 검색결과를 향상시키는 방법을 제안한다. 본 논문에서는 데이터 집합을 분류해내기 위해서 각 속성간의 거리에 가중치를 적용하는 k-nearest neighbor 분류법을 사용하였고, 속성 가중치를 동적으로 변화시키는 규칙을 추출하기 위한 방법으로는 결정 트리 생성에 의한 규칙(decision rule) 생성 방법을 적용하였다. 검색결과 향상을 보이기 위한 실험으로써 온라인 커플매칭(online couple-matching) 시스템의 핵심부분을 구현하고 이를 적용하였다.

Key words: 속성 가중치(attribute weighting), 커플매칭(couple-matching), k-nearest neighbor, 결정 규칙(decision rule)

1. 서 론

정보화 사회에서 효과적인 지식 추출의 도구로서 전 산업분야에 적용되고 있는 데이터 마이닝(Data Mining) 기술은 자동화되고 지능을 갖춘 데이터베이스 분석기법으로 90년대 초반부터 지식발견(KDD : Knowledge Discovery in Database), 정보발견(Information Discovery), 정보수확(Informat Harvesting) 등의 이름으로 소개되어 왔다[Chen97]

* 숭실대학교 컴퓨터학과 석사과정
** 숭실대학교 컴퓨터학부 부교수

그러나 현재 데이터 마이닝은 일반적으로 지식발견의 과정 중 대용량의 데이터베이스의 데이터로부터 패턴인식, 통계적 기법, 인공지능 기법 등을 이용하여 숨겨져 있는 데이터간의 상호 관련성, 패턴, 경향 등을 추출하는 것으로 정의한다. 또한 상거래 분석, 판매전략 수립, 수요예측, 고객관리 등의 의사결정에 활용하는 도구로서 정의하며, 다양한 종류의 새로운 기법들에 대한 많은 연구가 진행되고 있다. 데이터 마이닝의 종류로는 연관규칙 발견(Association Rule Discovery), 연관성 분석(Link Analysis), 분류(Classification), 추정(Estimation), 예측(Prediction), 군집화(Clustering) 등이 존재한다[Chen97][Agrawal93].

최근의 인터넷 사용 현황을 보면 크게 정보검색

과 엔터테인먼트의 두 부류로 분류해 볼 수 있다. 지금까지 정보검색 분야에 관련하여 지식추출과 이를 활용하기 위해 여러 가지 데이터 마이닝 기법들을 적용해왔다. 본 연구는 인터넷 활용의 한 부분인 엔터테인먼트 분야를 적용대상으로 하며, 데이터 마이닝 기법과 인공지능 기법을 함께 적용하여 사용자의 입력정보를 바탕으로 지식을 발견하고 기존의 데이터들의 중요도를 조절하는 에이전트의 지식으로 활용하려 한다. 또한 본 연구는 속성의 가중치 변경을 통한 이상형 정보 검색 결과의 질을 향상시키는 것을 목적으로 하고 있다. 이를 위해서 사용자 적합성 평가(user relevance feedback)를 통한 가중치 부여 방식에 대해 연구하였으며, 보다 효율적인 근사 가중치 적용법(approximate weighting method)을 제안하고 있다.

본 논문의 2장에서는 기존의 데이터 분류법에서 속성의 가중치를 부여하는 방식에 대한 연구를 통해 그 문제점을 알아보고, 3장에서는 본 논문을 통하여 구축하고자 하는 매칭 에이전트의 개략적인 구조와 기능에 대하여 기술한다. 4장에서는 이상형 정보 검색의 결과와 사용자 적합성 평가를 통하여 기존의 근사 가중치 적용법을 개선함으로써 사용자에게 보다 적합한 결과를 산출하는 접근방법에 대하여 설명하였으며, 5장에서는 실제 사용자들의 매칭 입력 데이터를 대상으로 동적인 가중치 조절을 적용한 사례 실험을 통하여 그 효용성을 고찰해 보기로 한다.

2. 관련 연구

2-1. k-nearest neighbor

k-NN(k-Nearest Neighbor)은 기본적인 분류 알고리즘으로 임의의 점과 가까운 k개의 점을 이용하여 임의의 점의 이산적인 값 또는 연속적인 변수를 산정하는 방법이다.

$$A = < a_1(x), a_2(x), ..., a_n(x) >$$
$$W = < w_1, w_2, ..., w_n >$$
$$d(x_i, x_j) \equiv \sqrt{\sum_{r=1}^{n} w_r (a_r(x_i) - a_r(x_j))^2}$$

k-NN의 연속변수 산정 방법은 임의의 점과 가까운 k개의 점을 평균하여 연속적인 변수를 산정하는 방법과 거리에 따라 가중치를 다르게 하여 연속변수를 산정하는 방법 등이 존재한다[Mitchell97]. 그러나 k-NN은 단순히 점과 점 사이의 거리를 이용하여 연속변수를 산정하기 때문에 속성(attribute)의 중요도가 속성 값들의 간격에 의해 결정된다는 문제점이 있다.

2-2. 근사 가중치 적용법

근사 가중치 적용법은 다차원의 속성을 갖는 데이터 모델에서 속성의 값에 가중치를 적용하는 문제에 유용하게 쓰인다. 대표적인 3가지 방식은 equal weights, ROC(Rank-oder-centriod) weights Rank-sum weights를 들 수 있다. equal weights는 가중치의 상대성을 무시하고 일률적으로 가중치를 부여하는 방식이다. ROC weights는 속성의 중요 순위에 따라 가중치를 부여하는 방식이며, Rank-sum weights는 ROC에 가중치의 순위를 적용하는 방식이다. 일반적으로 ROC weights와 Rank-sum weights의 성능이나 효용성은 크게 차이가 나지 않는 것으로 알려져 있다. 따라서, 사용하는 데이터 모델에 따라 적절히 선택할 수 있다. ROC weights 방식은 다음과 같다.

$$w_{(i)} = \frac{1}{m} \sum_{k=i}^{m} \frac{1}{k}, \quad i = 1,2,...,m$$

여기서, $w_{(1)} \geq w_{(2)} \geq ... w_{(m)} \geq 0, \quad \sum_i w_{(i)} = 1$ 이며, m은 속성의 개수이다[Jianmin93].

3. 매칭 에이전트의 구조

본 논문에서 모델로서 다루는 매칭 에이전트는 그림 1에서 보는 바와 같이 사용자의 이상형 검색 질의어를 통하여 이상형에 가까운 결과를 순위별로 얻게되고 사용자 적합성 평가를 통하여 가중치 부여 정책을 결정하는 피드백 학습을 수행하며, 피드백 학습으로부터 얻는 결정규칙과 검색결과에서 추출되는 추가적인 가중치 요소를 적용하여 가중치를 동적으로 조절함으로써 사용자에게 보다 적합한 검색결과를 산출하는 것으로 구성되어 있다.

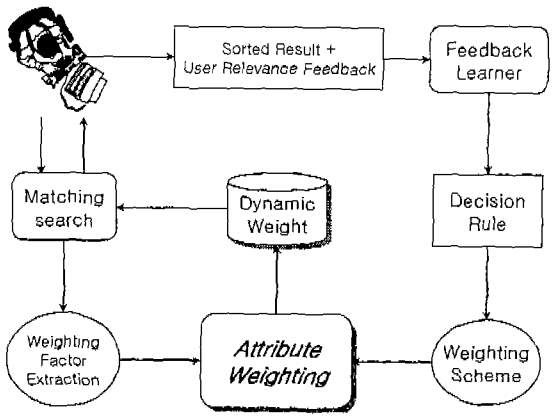

그림 1 매칭 에이전트의 구조

매칭 에이전트는 크게 3부분의 하위 기능 모듈로 구성된다.

3-1. k-NN 분류법을 이용한 이상형 검색

이상형 검색을 위한 사용자 인터페이스 모듈이다.

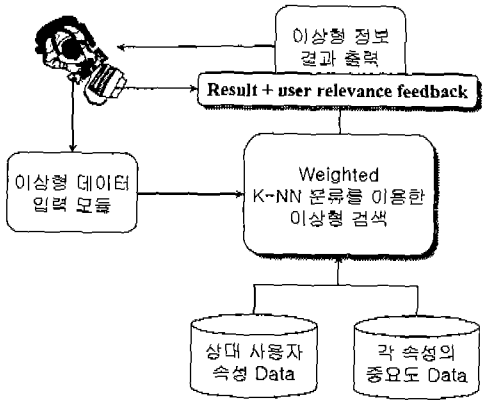

그림 2 이상형 검색 모듈

사용자로부터 이상형에 대한 각 속성값과 그에 대한 중요도를 입력받으며, 사용자와의 통합적인 상호작용 전반에 걸친 기능과 매칭순위로 정렬된 검색결과를 제공함으로써 사용자 적합성 평가를 가능하게 한다.

3-2. 피드백 학습

사용자의 적합성 평가를 기반으로 피드백 학습을 수행하는 모듈로서, 귀납적 기계학습을 통해 생성되는 결정규칙을 이용하여 가중치 부여 정책을 결정한다.

3-3. 동적 가중치 적용

사용자의 질의와 대응하는 속성값의 가중치를 동적으로 조절하는 모듈로서, 가중치 부여 정책에 따라 검색결과에서 추출된 가중치 요소를 적용하여 새로운 가중치를 재계산하고 이를 저장하는 역할을 수행한다.

4. 동적 가중치 조절 방식

본 논문에서는 동적인 가중치 조절 방식을 구현하기 위해서 크게 2가지를 주안점으로 두고 있다. 첫째, 속성의 불연속적인 값의 전처리 과정으로서, 기존의 k-NN 분류법의 거리(distance) 계산과 가중치 부여 방식은 수치적이고 연속적인 값들만을 처리하는데 비해 본 논문은 비수치적이고 불연속적인 값들을 함께 처리하는 방식을 취하고 있다. 둘째, 기존의 근사 가중치 적용법을 개선하기 위한 모듈로서, 피드백 학습을 통한 가중치 부여 정책 결정 과정과 검색 결과로부터 얻어지는 결과의 순위차와 거리차를 적용하는 방식을 취하고 있다.

4-1. 불연속적인 속성 값

실세계의 데이터는 수치적이고 연속적인 값과 비수치적이고 불연속적인 값이 존재한다. 본 논문에서 취하는 k-NN 분류법의 거리 계산과 가중치 적용에 있어서 비수치적이고 불연속적인 값을 직접 입력으로써 적용하는 것은 불가능하다. 따라서 이를 위한 전처리 과정이 불가피하며, 불연속적인 변수를 등급을 갖는 변수와 다중 값을 갖는 변수로 분류하여 처리하고 있다.

첫째, 등급을 갖는 불연속적인 변수는 수치적인 값으로의 매핑 테이블을 이용하여 수치적인 값으로 변환한다. 그림 3에서 *character*라는 속성은 6등급을 갖는 불연속적인 값으로 볼 수 있다. 따라서 매핑 테이블을 통해 수치적이고 연속적인 값으로 변환할 수 있는데, 다른 연속적인 속성 변수와 데이터의 범위가 다를 수 있음을 고려하여 정규화(normalization) 시켜야 한다. 이 경우는 C4.5 학습에서는 변환될 필요가 없으며 단지 거리 계산과 가중치 적용시에만 변환된다.

둘째, 그림 3에서 *hobby*와 같이 다중 값을 갖는 불연속적인 변수는 속성값의 중요순위에 따라 상대

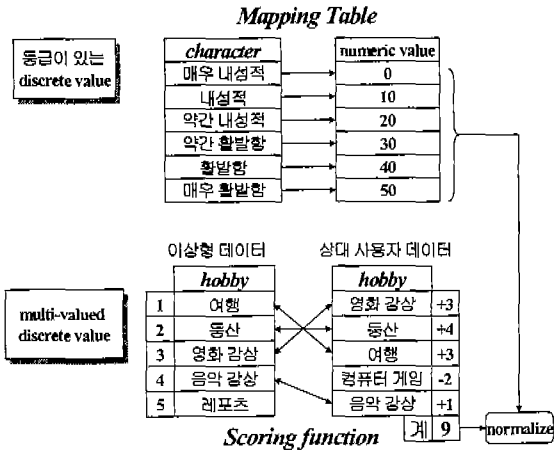

그림 3 불연속적인 값의 처리

사용자의 데이터와 매핑되는 정확도에 따라 점수를 부여하는 방식을 취하며, 이 점수는 정규화를 통하여 그 속성의 연속적인 값으로 변환된다. 이 경우는 입력 초기에 연속적인 값으로의 변환이 필요하다.

위의 두 가지 방식에서 사용되는 정규화 방식은 다음과 같다[Jianmin93].

$$v_i(x_i) = \frac{(x_i - x_{i,\min})}{(x_{i,\max} - x_{i,\min})}$$

여기서, $v_i(x_i)$는 $0 \le v_i(x_i) \le 1$의 값을 갖는다.

4-2. 동적 가중치 조절

이상형 정보의 속성값의 입력을 질의어로 하여 초기 검색 결과가 산출되면 사용자는 그 결과에 대하여 적합성을 평가(user relevance feedback)한다.

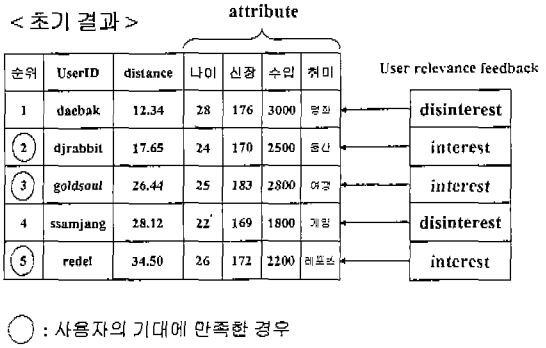

그림 4 초기결과와 사용자 적합성 평가

초기 검색결과에서 사용자에게 적합한 경우는 사용자의 기대에 만족한 정보가 결과 목록에 포함되어 있으며 높은 순위를 차지하는 경우이다. 반면, 사용자에게 적합하지 못한 경우는 사용자가 원하지 않는 이상형 정보가 결과 목록에 포함되거나 사용자가 원하는 정보라 할지라도 그 순위가 적절하지 못한 경우이다.

초기 결과가 사용자에게 적합하지 못한 경우에는 그림 4에서와 같이 사용자 적합성 평가를 명시적으로 표현하여 준다. 그림에서 보는 바와 같이 초기 결과에 나타난 속성들의 값과 사용자 적합성 평가를 통해서 학습 데이터를 얻을 수 있다. 즉, 명시적인 사용자 적합성 평가는 클래스(class)가 된다. 이렇게 얻어진 학습 데이터를 가지고 본 논문에서는 C4.5[Quinlan92]를 이용한 피드백 학습을 통하여 결정규칙을 생성하였다. 피드백 학습을 하는 목적은 가중치 부여 정책을 결정하기 위함이다. 가중치 부여 정책이란, 사용자가 적합하다고 평가한 데이터 집합은 궁극적으로 높은 순위를 차지해야 하므로 k-NN으로부터 계산되는 거리값을 낮춰주어야 한다. 이를 위해 속성의 가중치를 충분히 높여주어야 한다. 반대의 경우는 물론 사용자가 적합하지 않다고 평가한 경우이다. 다음은 C4.5를 이용한 결정규칙의 한 예이다.

```
Rule 1:
  age <= 27
  height > 162
  income > 2000
  -> class interest

Rule 2:
  character = shy
  income <= 1800
  -> class disinterest
```

Rule 1에 나타나는 조건들을 만족하는 데이터 집합들은 가중치 부여 정책에 따라서 가중치를 충분히 높여 주어야 한다. 반면 Rule 2에 나타나는 조건들에 만족하는 데이터 집합들은 가중치를 충분히 낮춰 주어야 한다.

일단 가중치 부여 정책이 결정되면, 효율적인 근사 가중치 적용을 위해서 초기 결과에서 결과의 순위차와 거리차를 추출한다. 결과의 순위차란, 현재 데이터 집합의 순위과 목적순위와의 차이를 말한다. 목적순위는 사용자가 적합하다고 평가한 데이터 집합일 경우는 1이 되고, 그 반대의 경우는 최하위 순위가 된다. 또한 결과의 거리차란, 현재 데

이터 집합의 거리(distance)와 목적 순위에 있는 결과의 거리와의 차이를 말한다.

추출한 두 가지 요소를 적용하여 본 논문에서 제안하는 근사 가중치 적용 방식은 다음과 같이 표현할 수 있다.

$$w_{(i)} = ROC(w_{(i)}) \times | r_c - r_{dest} | \times | d_c - d_{dest} |$$

위의 수식에서 ROC($w_{(i)}$)는 2장에서 언급된 ROC weights 방식으로 얻어지는 가중치이다. r_c, r_{dest}는 각각 현재 데이터 집합의 순위와 목적순위를 나타내는 것이며, d_c, d_{dest}는 각각 현재 데이터 집합의 거리와 목적순위에 있는 데이터 집합의 거리를 나타낸다. 위의 수식에서 얻어지는 새로운 가중치는 동적 가중치 데이터베이스에 저장되며, 이를 통해 순환적인(iterative) 시스템을 가능하게 한다.

5. 실 험

본 실험의 목적은 가중치의 동적인 변경과 적용을 통하여 매칭 에이전트의 결과의 질을 어떻게 향상시키는가를 평가하기 위한 것이다. 표 1은 사용자가 이상형 검색을 하기 위한 입력 데이터 형식이다.

속 성	속성값 범위
성별	남자, 여자
나이	10-60
지역	주요 시/도
성격	매우 내성적, 내성적, 약간 내성적, 활발함, 매우 활발함
취미	여행, 등산, 영화, 음악, 레포츠, 게임... 중 중요순위로 3가지
키	150-200
체형	0-50
헤어스타일	짧은 머리, 긴 생머리, 스포츠...
외모 중요도	0-50
혈액형	A, B, AB, O
학력	중졸, 고졸, 대졸, 석사, 박사
수입	0-5000
종교	기독교, 불교, 천주교, 기타
흡연	비흡연, 흡연
주량	0-50
원하는 관계	펜팔, 친구, 결혼, 연인...

표 1 입력 데이터 형식

실험의 방법은 본 논문에서 제안하는 동적 가중치 조절의 방식을 이용한 경우와 기존의 근사 가중치 방식만 적용한 경우를 비교 실험하였다. 실험의 평가 방식은 사용자가 평가한 데이터 집합의 순위가 사용자가 원하는 형태로 올바르게 변동되는가를 검사하여 정확도(%)로 나타내었다. 표 2는 비교실험 10회에 걸쳐 평균적인 결과를 나타내고 있다.

정확도 평가	기존의 근사 가중치(ROC weights)	제안하는 동적 가중치 적용방식
	63.4 %	81.7 %

표 2 비교 실험 결과

실험의 결과에서 보는 바와 같이 기존의 근사 가중치를 적용하는 방식보다 본 논문에서 제안하는 방식의 정확도가 평균적으로 18.3%가 높게 나타났다.

6. 결론 및 향후과제

본 논문에서 제안하는 사용자 적합성 평가를 통한 동적 가중치 조절 방식은 기존의 근사 가중치 적용법 보다는 사용자에게 적합한 결과를 제시하였다. 그러나 C4.5는 결정트리를 생성할 때, 데이터의 분류(classification)만을 정확히 하려고 하기 때문 결정규칙의 오차가 클 수 있다. 따라서 C4.5를 이용하는 방식이 아닌 보다 효율적인 특징 추출(feature selection)을 통하여 실험해 볼 필요가 으며, 동적으로 가중치를 변경하는 기존의 알고리즘과의 비교 실험도 필수적이라고 본다. 또한 본 실험에서는 실험자가 임의로 데이터의 속성 중 실험에 영향을 줄 속성들을 선택하였지만, 속성 선택을 자동화할 수 있는 기법에 대한 연구도 필요할 것이다.

참 고 문 헌

[Chen97] S. Chen, J. Han and P. Yu, *"Data Mining : An Overview form Database Perspective"*, IEEE Trans. on Knowledge and Data Engineering, 1997

[Agrawal93] Rakesh Agrawal, Tomasz Imielinski and Arun Swami, *"Data Mining : A Performance Perspective"*, IEEE Transactions

on Knowledge and Data Engineering, vol. 5, No. 6, December 1993, pp.914-925

[Mitchell97] T. M. Mitchell, *"Machine Learni* The McGraw-Hill Co., pp.230-248

[Quinlan92] J. Ross Quinlan, *"C4.5: PROGRAMS FOR MACHINE LEARNING"*, Morgan Kaufmann Publishers, 1992, pp.17-107

[Dean95] Tomas Dean, James Allen, and Yiannis Aloimonos, *"Artificial Intelligence Theory Practice"*, Cummings Publishing Co., 1995 pp.179-202

[Ginsberg92] Matt Ginsberg, *"Essentials Artificial Intelligence"*, Morgan Kaufmann P 1992, pp.313-318

[Chen91] S. Chen, C.F.N Cowan, and P.M. Grant, *"Orthogoanl Least Squares Learning fo Radial Basis Function Networks"*, IEEE Transactions on Neural Networks, 1991

[Lippmann87] Richard P. Lippmann, *"An Introduction to Computing with Neural Nets* IEEE ASSP MAGAZINE APRIL 1987

[Jianmin93] Jianmin Jia, Gregory W. Fischer, an James S. Dyer, *"Attribute Weighting Method and Decision Quality in the Presence of Response Error: A Simulation Study"*, Journal of Behavio Decision Making, May 1993