제20회 한국정보처리학회 추계학술발표대회 논문집 제10권 제2호 (2003. 11)

사례 기반 추론을 통한 전자상거래 상품 가시화 도구의 설계 및 구현

김주하*, 권기현*
*삼척대학교 정보통신공학과
e-mail: kimjuha96@hotmail.com

Design and Implementation of Data Visualization Tool using Case-Based Reasoning on Electronic Commerce

Ju-Ha Kim*, Ki-Hyeon Kweon*
* Dept of Info. & Communication Engineering, Samcheok Natl. Univ.

요 약

전자상거래 상의 방대한 데이터베이스의 자료 중에서 검색한 정보를 직관적으로 선택할 수 있도록 하기 위해서는 효율적인 검색 기능뿐만 아니라 검색된 결과의 표현 및 가시화에 대한 부분이 중요하다. 현재까지 검색 방법의 효율성에 대한 연구는 많이 진행되고 있으나 검색 결과의 가시화 방법에 대한 연구는 미미한 형편이다. 본 연구에서는 전자상거래를 위한 검색 결과를 유사도를 기준으로 가시화 시키는 데이터 가시화에 대한 연구를 하였다. 유사도는 유클리드 거리를 기준으로 Nearest Neighbor 방법을 사용하여 2차원 평면상에 상품을 가시화하도록 하는 전자상거래 상품 가시화 에이전트를 설계하고 구현한다.

1. 서론

최근 인터넷의 폭발적인 증가로 인해 사용자는 많은 양의 정보를 접하게 되었고 많은 양의 정보 중에서 자신이 원하는 정보만을 구하려는 요구가 발생하게 되었다. 따라서, 검색엔진을 사용하고 있으나 데이터베이스 내의 자료의 자료 양이 방대해 짐에 따라 불필요한 정보가 많이 포함되고 있다. 또한, 정보의 검색도 중요하지만 검색한 정보를 가시화하여 사용자에게 직관적인 판단을 할 수 있도록 하는 방법도 필요하게 되었다. 현재까지 검색 방법의 효율성에 대한 연구는 많이 진행[1][2][3]되고 있으나 검색 결과의 가시화 방법에 대한 연구는 미미한 형편이다.

따라서, 본 연구에서는 전자상거래에서 검색 결과를 가시화 시키는 데이터 가시화 방향에서 상품 검색 결과를 유사도에 따라 가시화 하는 연구를 제안한다. 유사도는 유클리드 거리를 기준으로 Nearest Neighbor 방법을 사용하여 2차원 평면상에 상품을 가시화한다.

본 논문은 2장에 관련 연구를 기술하고 있으며, 3장에서 IBL과 K-Nearest Neighbor 방법을 전자상거래 상의 상품 유사도에 사용하는지에 대해 설명하고, 4장에서는 사례기반추론에 기반한 가시화 에이전트 구현 내용에 대해서 설명하며 끝으로 결론 및 향후 연구 방향을 제시한다.

2. 관련연구

2.1 IBL

IBL(Instance Based Learning)은 저장된 사례에서 새로운 사례를 일반화시키는 방법이다. 저장된 사례는 새로운 사례가 추가될 때 가공되어 관계가 결정되어진다. 이 방법은 새로운 사례가 추가될 때까지 대기하게 되므로 지연 학습 방법(lazy learning)이라고 한다[9]. 이 방법은 검색 시스템에서 사용되어 기존의 사례에 대한 최적 매치, 인접 매치, 유사 매치 형태로 기존에 저장된 사례를 찾는데 사용된다. 이 방법을 전자상거래에 적용하면 상품을 검색할 때 찾고자하는 상품과 쇼핑몰에서 가지고 있는 상품 사이에 최적의 상품을 찾는데 적용 될 수 있다.

많이 사용되고 있는 IBL 방법에는 K-NN(K-Nearest Neighbour) 방법, Locally Weighted Regression, Radial Basis Function 방법이 있다. 본 연구에서는 이들 방법 중 K-NN(K-Nearest Neighbour)[8, 9] 방법을 사용하여 전자상거래 상의 상

품을 효율적으로 검색하고 가시화하는 에이전트를 설계하고 구현하고자 한다.

2.2 K-NN

K-NN(Nearest Neighbor) 방법은 IBL 방법중 가장 간단한 방법으로 사례를 n 차원 공간(R_n)에 있는 점으로 정의할 수 있다. K-NN 방법은 표준 유클리드 거리(R_n에 있는 대상간의 거리)를 사용하여 관계를 정의한다[8]. 예를 들어, x 라는 사례의 속성이 $< a_1(x), a_2(x), a_3(x), \ldots, a_n(x)>$와 같고 $a_r(x)$를 인스턴스 x의 r번째 속성이라고 하면 두 대상 x_i와 x_j의 거리는 식 (1)로 구해진다[7].

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{r=n} [a_r(x_i) - a_r(x_j)]^2} \quad (1)$$

이때 각 대상과 검색하고자하는 대상 x_q에 대한 거리를 구하여 가장 근접한 대상이 최적 매치가 된다.

거리 기반 가중치를 부여하여 nearest neighbor를 구하고자 하는 경우에는 대상의 각 속성에 대한 가중치를 식 (2)와 같이 구하여 사용한다[9].

$$w_i = \frac{1}{d(x_q, x_i)^2} \quad (2)$$

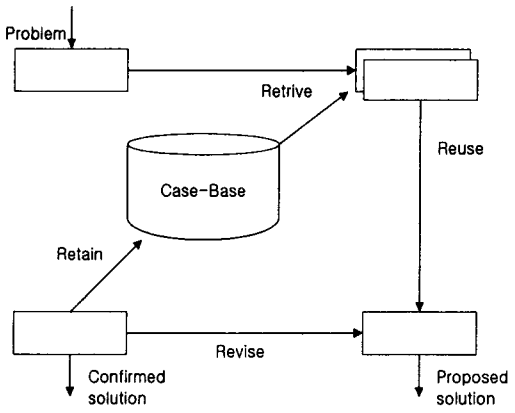
본 연구에서는 가중치를 식 (1)에 적용하여 사용하지 않고 사용자로부터 가중치를 입력받아 사용한다.

2.3 사례기반추론

사례기반추론(Case Based Reasoning)은 이전의 대상들에서 새로운 사례를 추론하기 위한 인공 지능적인 방법으로 지식 베이스를 기반으로 정한 기준에 맞는 유사한 사례를 추론한다. 이 기법은 새로운 문제가 발생하면 이미 경험한 지식 베이스에서 가장 유사한 사례를 검색하여 이전 사례를 재사용하는 적응 과정을 통해 새로운 문제를 해결하는 방식이다. 또한, 교정과정에서 새로운 문제의 해가 문제 해결에 적합한지를 검정하여 새로운 사례로 학습한다[6]. (그림 1)은 일반적인 사례기반추론 기법의 개념도 이다.

사례기반추론 기법에서 가장 유사한 사례를 찾는 방법으로 Nearest-Neighbor 기법과 귀납적 기법을 주로 사용한다. 아래 식 (3)은 Nearest-Neighbor 기법에서 입력문제 T와 사례 S에 대한 유사도를 계산하는 식이다. 식 (3)에서 T는 입력 문제, S는 학습된 사례, W_i는 T와 S의 각 속성에 대한 가중치로 정의된다[5].

$$Similarity(T, S) = \sum_{i=1}^{i=n} f(T_i, S_i) \times W_i \quad (3)$$



(그림 1) 사례기반추론 기법의 개념도

3. IBL 및 K-NN을 이용한 유사도 계산

전자상거래 상품 가시화 에이전트를 위한 속성 데이터에 대한 정규화, 가중치 부여, 유사도 계산 방법을 설명한다.

3.1 정규화

사례기반추론에 사용되는 원 데이터는 속성에 따라 크기 및 범위가 모두 상이하므로 그대로 사용하기에는 무리가 따르므로 정규화(normalization)을 통해 사용한다. 정규화는 식 (4)를 사용한다. 식 (4)에서 a_i는 대상의 속성을 SD는 표준편차를 a_m은 정규화된 속성 값을 의미한다.

$$a_m = \frac{a_i - \frac{\sum_{i=1}^{i=n}}{n}}{SD} \quad (4)$$

3.2 속성에 가중치 부여

각 대상의 속성에 가중치를 부여하여 사용한다. 가중치는 사용자의 입력에 의해 부여되며 가장 높음을 5, 가장 낮음을 1로 하여 사용한다.

<표 1> 가중치 테이블

가중치 의미	값
매우 낮음	1
낮음	2
보통	3
높음	4
매우 높음	5

3.3 유클리드 거리 계산

유클리드 거리 계산은 원 대상 자료에 식 (5)를 적용하여 구한다. 이때 대상들의 각 속성의 값이 가장 최소인 대상과 가장 큰 대상을 구해 이 두 대상간의 유클리드 거리를 최대 거리로 사용한다. 찾고자하는 대상과 기존 대상 모두 최대 거리(MAX_DISTANCE) 안에 있게 된다.

$$d(x_i,x_j)=\sqrt{\sum_{r=1}^{r=n}[a_r(x_i)-a_r(x_j)]^2} \qquad (5)$$

3.4 유사도 계산

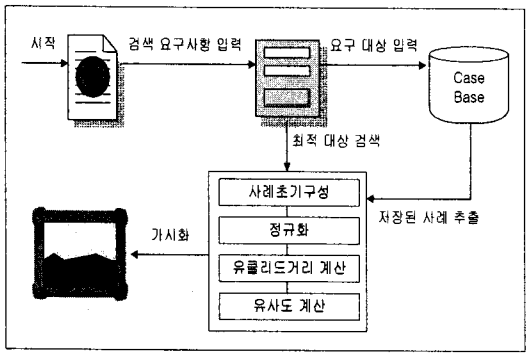
유사도는 유클리드 거리를 최대 거리로 나눈 값을 비율로 나타낸다. 찾고자하는 대상과 일치하는 대상의 유사도는 100%로 비율이 낮을수록 유사도가 떨어짐을 의미한다. 식 1에서 d_i는 임의 대상의 유클리드 거리, d_{max}는 최대 유클리드 거리, x_i는 임의 대상, x_q는 찾고자하는 대상을 의미한다.

$$Similarity(x_i,x_q)=\left(1-\left(\frac{d_i}{d_{max}}\right)\right)\times 100 \qquad (6)$$

4. 사례기반추론 가시화 에이전트 구현

4.1 프로토타입 빌더 제한 사항

(그림 2)는 사례기반추론 가시화 에이전트의 구조로 웹 페이지에 방문하여 질의 상품 정보와 가중치를 입력하면 질의 대상을 데이터베이스에 입력하고 기존의 대상을 추출하여 가중치를 고려한 정규화, 유클리드 거리 계산, 유사도 계산을 통해 2차원 평면에 가시화하는 구조이다.



(그림 2) 전자상거래 상품 가시화 에이전트의 구조

4.2 구현 환경

구현 환경은 Apache 웹 서버에서 자바 애플릿을 클라이언트로 하고 MySQL 3.23.02 데이터베이스에 JDBC를 연동하여 구현하였다.

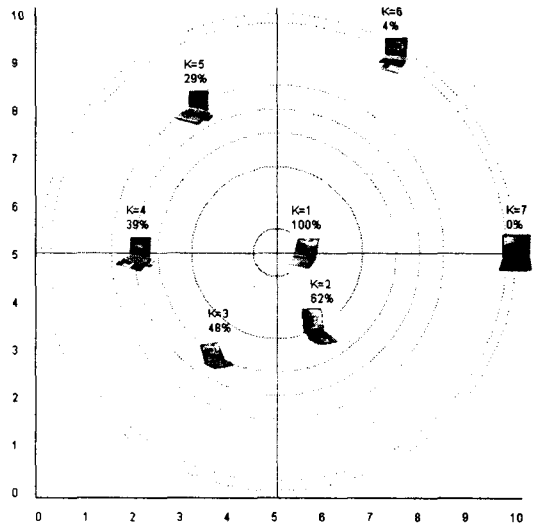
4.3 알고리즘

CBRSearch 클래스는 사례기반추론 가시화 에이전트의 핵심 클래스로 사용자 질의 대상과 데이터베이스의 대상을 사용하여 정규화, 유클리드 거리 계산, 최대 유클리드 범위 계산, 유사도 계산을 통해 가시화를 위한 데이터를 구하는 클래스이다. 이 클래스의 주요 흐름은 다음과 같다.

```
public class CBRSearch {
 public void init() {
 // DB에 저장된 대상과 질의 대상을 Vector로 작성
 // makeDataInit();

 // 가중치 적용한 makeDataInit()
 makeDataInit(weightvalue);

 // 속성 값이 최대인 대상과 최소인 대상을 구함
 getMaxMinInstance();

 // 속성 값이 최대, 최소인 대상을 벡터로 작성
 minInstanceVec = getVectorInstance(minInstanceArr);
 maxInstanceVec =
getVectorInstance(maxInstanceArr);

 // 정규화 수행
 composeNormalize();

 // 최소, 최대 대상에 대해서 정규화 수행
 minmaxInstanceNormalize();

 // 유클리드 디스턴스 구함
 euclideDistanceVec = calcEuclideDistance();

 // 모든 대상 간에서 최대 거리를 구함
 double LONG_DISTANCE =
        calcEuclideDistance(minNormalizedVec,
maxNormalizedVec);

 // 유클리드 디스턴스를 정렬함
 sortedDistance =
                calcSortedDistance(euclideDistanceVec);

 // 최대 거리값을 사용하여 유사도를 계산
 similarity =
     calcSimilarity(sortedDistance, LONG_DISTANCE);

 // 그래픽 좌표로 나타내기 위해 거리를 크게 조정함
 scaledDistance = calcScaledDistance(sortedDistance,
LONG_DISTANCE);
 // 가시화
 repaint();
 }
}
```

4.4 구현 결과

구현 결과는 (그림 3)과 같으며 중심에 k가 1인 상품이 유사도 100%인 최적 매치이며 동심원이 멀어질수록 유사도가 떨어지는 것을 표현한 것이다. 화면에서 x또는 y축의 방향은 의미를 가지지 않으며 동심원에서 떨어진 거리만이 의미를 가진다. 사용자가 다른 상품의 그림을 클릭하면 선택된 상품을 기준으로 새로운 그림이 그려지거나 해당 상품에 대한 상세한 정보를 볼 수 있다.



(그림 3) 가시화 에이전트 사례기반추론 결과

5. 결론 및 연구 방향

수많은 검색 대상에서 원하는 대상을 빠르고 효율적으로 검색하는 연구가 많이 진행되어 왔으나 데이터 가시화에 대한 연구는 미미한 형편이다.

본 논문에서는 대상의 수치 데이터에 대해 유클리드 거리를 기준으로 Nearest Neighbor 방법을 사용하여 질의 대상과 저장된 대상간의 유사도를 구하였으며 이 값을 전자상거래의 상품에 적용하여 2차원 평면에 물품을 표현하는 데이터 가시화 에이전트를 구현하였다.

K-Nearest Neighbor 문제는 대상이 많은 경우에 사전에 가중치를 계산하여 저장하는데 어려움이 있고, 계산깊이를 결정하는 문제가 있으므로 이 부분에 대한 연구가 진행되는 것이 요구된다.

참고문헌

[1] 성백균, 김상회, 박덕원, "전자상거래를 위한 사례기반추론의 판매지원 에이전트", 한국정보처리학회 논문지, 제7권 제5호, pp.1649-1656, 2000. 5

[2] 백혜정, 장영택, "기계학습 기반 적응형 전자상거래 에이전트 설계", 한국정보처리학회 논문지, 제9-B권 제6호, pp.775-782, 2002. 12

[3] 김영설, 김병천, 윤병주, "개선된 추천시스템을 이용한 전자상거래 시스템 설계 및 구현", 한국정보처리학회 논문지, 제9-D권 제2호, pp.329-336, 2002. 4

[4] 황병연, 박성철, "전자상거래를 위한 정책지향 매칭 에이전트 시스템의 설계 및 구현", 한국정보처리학회 논문지, 제8-D권 제5호, pp.623-630, 2001. 10

[5] 김영지, "사례기반추론 기법을 이용한 개인화된 추천시스템 설계 및 구현", 한국정보처리학회 논문지, 제9-D권 제6호, pp.1009-1014, 2002. 12

[6] R. Schank, "Dymanic Memory: A Theory of Learning in computer and People," Cambridge University Press, New York, 1982.

[7] J.B. Schafer, J. Konstan, and J. Riedl, "Recommender Systems in E-Commerce, " Proceedings of ACM Conference on Electronic Commerce, November 3-5, 1999.

[8] J. Yang et al., "An inter-pattern distance-based constructive learning algorithm," Intelligent Data Analysis, ELSEVIER, 1999, pp.55-73.

[9] http://www.developer.com/java/other/article.php/1491651

[10] A. Moukas, R. Guttman, and P. Maes, "Agent-Mediated Electronic Commerce : An MIT Media Laboratory Perspective," Proceeding of Int. Conf. On Electronic Commerce, Seoul Korea, pp.9-15, 1998.

[11] T. M. Mitchell, "Machine Learning," McGraw Hill, 1997.

[12] G. Karypis, "Evaluation og Item-Based Top-N Recommendation Algorithms, "Technical Report CS-TR-00-46, Computer Science Dept., University of Minnesota, 2000.

[13] http://sourceforge.net/projects/selectionengine