퍼지 및 지능시스템학회 논문지 2006, Vol. 16, No. 2, pp. 156-163

새로 출시되는 품목들을 위한 단어 기반의 사용자 선호도 예측 기법

A Prediction System of User Preferences for Newly Released Items Based on Words

최윤석* · 문병로**

Yoon-Seok Choi and Byung-Ro Moon

* 서울대학교 컴퓨터공학부

** 서울대학교 컴퓨터공학부

요 약

협동적 여과(CF) 시스템은 구현의 용이성과 뛰어난 성능으로 널리 활용되고 있다. 그러나 이 시스템은 데이터 희소성, 신상품 추천 불가, 추천 근거에 대한 설명 부족 등의 문제점을 포함하고 있어 이를 해결하기 위한 많은 연구가 진행되었다. 데이터 희소성 문제는 데이터의 누적에 따라 해결될 수 있지만, 협동적 여과 기법의 특성상 새로이 출시되는 품목에 대한 추천이 불가능하다. 이를 해결하기 위해 내용 기반(CB) 기법을 같이 사용하는 연구들이 제안되었다. 또한 협동적 여과 시스템은 추천 과정에 있어 추천 근거에 대한 설명을 제공하지 않는다. 본 연구에서는 추천에 대한 설명 기능을 포함하고 있는 선호 단어를 활용한 내용기반 예측 시스템을 제안한다. 이 시스템은 새로이 출시되는 영화에 대해 사용자의 영화에 대한 평가 정보를 예측하며, 추천의 근거가 되는 선호 단어를 제시한다. 또한 기존의 내용기반 예측 시스템에서 일어나는 속성 비매칭 문제로 인한 성능 저하를 막기 위해 기호 네트워크를 활용한 성능 개선 방법을 제안한다. 성능 비교를 위해 EachMovie 데이터베이스와 IMDb 사의 영화 홍보 데이터를 사용하였다.

Abstract

CF systems are widely used in recommendation due to the easy implementation and the outstanding performance. They have several problems such as the sparsity problem, the first-rater problem, and recommending explanation. Many studies are suggested to resolve these problems. While the influence of the sparsity problem lessens as the users' data are accumulated, but the first-rater problem is originated from the CF systems and there are a number of researches to overcome the disadvantages of CF systems based on the content-based methods. Also CF systems are black boxes, providing no explanation of working of the recommendation. In this paper we present a content-based prediction system based on the preference words, which exposes the reasoning behind a recommendation. Our system predicts user's rating of a new movie and we suggest a semiotic network-based method to solve the mismatching problem between the items. For experimental comparison, we used EachMovie and IMDb dataset.

Key words : 추천 시스템, 선호도 예측, 신상품 추천문제, 추천 설명, 속성 비매칭 문제

1. 서 론

사용자가 원하는 정보나 품목을 추천하는 작업은 정보 검색이나 전자 상거래 분야에서 상당히 중요하다. 광고 전문가나 정보 분류 전문가에 의존하던 추천 작업은 정보의 범람으로 인해 보다 빠르고 효과적인 기술이 필요하게 되었다. 거대한 데이터에서 정보를 추출하고 결과 예측을 위한 규칙이나 패턴 발견을 위해 데이터 마이닝 기법은 많은 역할을 수행하였다 [1].

협동적 여과 (Collaborative Filtering: CF) 시스템은 사용자 선호도 예측 시스템에서 가장 많이 활용되는 기법이다. CF 시스템은 먼저 사용자가 접한 품목에 대한 반응 정보를 수집하며 이 정보를 기반으로 다른 사용자들과의 유사도를 계산한다. 계산된 유사도를 기반으로 대상 사용자와 유사도가 높은 사람들의 집단, 즉 이웃 집단을 설정하여 이웃 집단의 구성원이 특정 품목에 보인 선호도를 분석함으로써 대상 사용자의 특정 품목에 대한 선호도를 예측한다 [2].

내용기반 (Content-based: CB) 시스템은 품목간의 유사성 분석이나 사용자 프로파일과 품목간의 연관성 분석을 기반으로 한다. 사용자가 과거에 접하였던 영화, 책, 음악 등의 품목의 내용을 분석하여 이와 유사한 품목을 추천하거나 품목의 내용을 기반으로 사용자의 프로파일을 생성하고 사용자의 프로파일과 연관성이 높은 품목을 추천하는 방식이다. 품목간의 유사도를 직접적으로 활용하는 방법 [3]과 사용자가 과거에 접한 품목의 내용 분석을 통하여 사용자의 프로파일을 구축하고 대상 품목과 프로파일과의 연관도를 계산하여 대상 품목에 대한 선호도를 계산하는 방법이 있다 [4, 5, 6, 7].

CF 시스템은 사용자나 품목의 속성에 대한 정보에 대한 심도 있는 고찰 없이도 적용할 수 있어 많이 사용되고 있으나 데이터 희소성과 신상품 문제(the first-rater problem) 등의 단점을 갖고 있다 [8]. 데이터 희소성은 아주 많은 품목

접수일자 : 2005년 12월 23일
완료일자 : 2006년 4월 11일

수에 비해 사용자가 실제 구입하거나 평가하는 품목의 수가 상대적으로 적기 때문에 비롯된다. 이로 인해 사용자간의 공통 품목을 기반으로 하는 사용자 유사성 분석이 제대로 이루지지 않게 된다. 또한 새로이 출시하는 품목은 접한 사용자가 없기 때문에 근본적으로 사용자의 선호도 예측이 불가능하다.

CB 시스템은 사용자 프로파일과 품목 또는 품목간의 공통 속성을 결정하여 공통 속성을 기반으로 평가 값을 예측한다. 하지만, 품목간의 공통 속성이 없거나 적은 경우에 유사도 측정의 정확도가 떨어지게 되는데 이를 속성 비매칭 문제(the mismatching problem)라 한다. 이러한 비매칭 문제는 정보 검색 분야에서 많이 나타나는 문제로 입력 질의의 단어가 검색 대상 문서에 존재하지 않거나 단어가 중의적인 의미를 함유하는 경우로 검색 품질을 떨어뜨리게 된다 [9].

본 논문에서는 CF 시스템의 단점인 새로운 품목에 대한 추천 불가 문제를 해결하기 위해 사용자가 과거에 접한 품목들의 속성(단어)을 이용하여 사용자의 선호도를 예측하는 CB 시스템을 제안하였다. 또한 CB 시스템에서 발생하는 속성 비매칭 문제에 대한 개선 방법으로 기호 네트워크(Semiotic Network) 기반의 확장 문서 유사도 시스템을 제안하였다.

이 논문은 다음과 같이 구성된다. 2절에서는 예측 기법의 기존 연구들을 설명하고, 3절에서는 본 논문에서 제안하는 선호 단어 기반 예측 시스템, 기호 네트워크를 활용한 비매칭 해결 방법을 설명하고 문서 유사도 시스템, 단순 베이지안 분류기, 선호 단어 기반 시스템과 확장 문서 유사도 시스템을 구축한다. 4절에서는 제안된 시스템을 기존의 방법과 실험을 통해 비교하고, 5절에서 결론을 맺는다.

2. 배 경

2.1 CF 시스템

사용자 기반의 CF 시스템은 Goldberg 등에 의해 Tapestry라는 전자 메일 여과시스템에 처음 언급되었다 [2]. CF 시스템은 사용자의 과거 행동 기록을 기반으로 사용자간의 연관성 고찰에 근거한다. 예를 들면 대상 사용자 A의 품목 x에 대한 선호도를 예측하고자 할 때 기존 사용자 중에서 A와 가장 유사한 성향을 보인 사용자들을 선별하여 이웃 집단(U)을 구성한다. U에 속한 사용자들이 품목 x에 보인 선호도를 분석하여 사용자 A의 품목 x에 대한 선호도를 예측한다. 이러한 사용자 기반 CF 시스템은 많은 연구가 진행되었으며 많은 기업에 의해 상용화가 되었다. CF 시스템을 사용한 대표적인 제품으로는 GroupLens [11], Video Recommender [12], Ringo [13], 그리고 LikeMinds [14] 등이 있다. CF 시스템은 품목의 속성 분석이 필요 없다는 점과 CB 시스템과 달리 사용자의 속성 정의에 제약이 없으며 이웃 집단에 속한 다양한 사용자의 정보를 활용함으로써 다양한 품목을 접할 좋은 기회를 제공하기도 한다.

Jonathan 등은 CF 시스템에서 발생하는 오류를 모델/프로세스 오류와 데이터 오류의 두 가지로 구분하였다 [15]. 그 중 데이터 오류에 의한 문제는 데이터 불충분으로 이는 데이터의 누락이나 평가가 적은 것을 의미하는 것으로 이를 데이터 희소성이라 한다. 두 번째 문제는 신상품 평가 문제이다 [16]. 새로운 품목에 대해서는 이웃집단에 속한 사용자들의 평가 데이터가 존재하지 않으므로 근본적으로 추천이 불가능하다.

2.2 CB 시스템

CB 시스템은 품목간의 유사성 분석이나 사용자 프로파일과 품목간의 연관성을 활용하는 방법으로 문서 검색 부분에서 먼저 활용되었다. 문서 내의 주제어나 단어의 단순 패턴 비교에서 시작되어 문서의 통계적, 구문적 분석을 통한 문서 분류 기법으로 발달하였으며 문서 내에 존재하는 단어와 구문을 추출하여 단어와 단어, 구문과 구문과의 관계 분석을 통하여 문서간의 유사도를 계산하기도 한다 [17].

사용자의 프로파일 기반 CB 시스템에서 주요 작업은 사용자 프로파일을 추출/정의하는 작업과 생성된 프로파일을 기반으로 품목에 대한 선호도를 계산하는 작업이다. 프로파일을 생성하는 가장 손쉬운 방법은 사용자로부터 상세한 정보를 얻는 일이지만, 대부분의 사용자는 정보를 제공하는 데에 적극적이지 않다. 따라서 사용자와 관련 있던 품목의 분석을 통해 자동으로 생성하는 방법이 주로 사용된다. 대표적인 시스템으로는 INFOSCOPE [18], NewsWeeder [4], InfoFinder [5], LIBRA [6], Syskill and Webert [7] 등이 있다.

CB 시스템의 문제점은 품목의 속성 추출과 사용자 프로파일 생성이 어렵다는 점이다. 책이나 웹 페이지 등의 문서 기반 시스템에서는 문서 분석을 사용할 수 있지만, 음악이나 영화, 의료, 식품 등은 그 본연의 속성을 추출하는 것이 쉽지 않기 때문이다 [20]. 또한 추출된 속성을 분석하여 어떠한 속성이 사용자의 선호도 결정에 기여도가 높은지를 결정하여 기여도가 높은 속성으로 프로파일을 생성해야 한다. 또한 CF 시스템의 데이터 희소성 문제와 유사한 속성 비매칭 문제가 있다. 품목간의 속성과 속성, 사용자의 프로파일과 품목 속성과의 공통 속성이 없거나 적은 경우로써 품목간의 유사도를 계산하지 못하거나 그 정확도가 떨어지게 된다.

2.3 기호 네트워크

기호 네트워크는 단어 간의 연관 관계를 그래프로 표시한 것이다. 그래프의 노드에는 단어가 할당되며 노드를 연결하는 연결선은 두 노드 간의 관계를 표현한다. 연결강도는 두 단어가 같은 문서에서 공통 출현하는 빈도를 분석함으로써 계산한다.

자연 언어 처리나 웹 검색 등의 분야에서 문서를 구성하는 단어간의 연관성을 네트워크로 표현하여 유사 단어 등을 검색하거나 검색을 확장하는 방법으로 널리 사용되었다 [21]. Lee 등은 단어를 하나의 기호로 간주한 기호 네트워크(Semiotic network)를 기반으로 구글 검색의 만족도를 10% 가량 향상시켰다 [22]. Xu 등은 문서 검색 분야에서 단어 유사성을 기반으로 사용자의 질의를 확장하여 검색의 효율을 20% 이상 향상시켰다 [23].

3. 선호 단어 집합 기반 예측 시스템

3.1 사용자 프로파일로서의 선호 단어 집합의 구축

CB 시스템 구축을 위해 먼저 품목의 속성 분석과 사용자의 선호 성향 분석이 필요하다. 먼저 품목 속성 분석은 품목을 어떠한 요소를 사용하여 품목의 특성을 정의할 것인가라는 것이다. 그 다음은 사용자의 성향 분석으로 이는 사용자의 선호도에 영향을 미치는 요소를 추출하는 작업이다. 어떤 사용자는 영화 스타의 팬클럽에 가입하고 그 배우가 출연하는 영화를 빼놓지 않고 보려고 한다. 다른 사용자는 전쟁이

나 서부 영화 장르의 영화를 좋아한다. 전자의 사용자는 영화 배우가 영화를 선택할 때의 기준이 되며, 후자의 사용자는 장르가 기준이 된다. 그러나 일반적인 시스템 환경에서 사용자가 자신의 요구나 선호 성향을 직접적으로 밝히는 경우는 기대하기 힘들다.

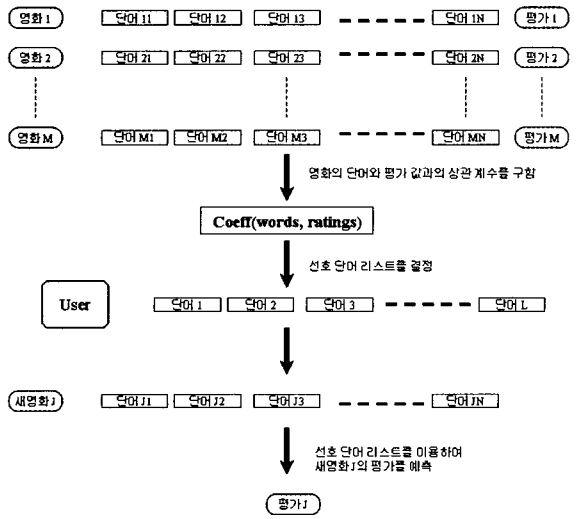

그림 1. 단어 기반의 새 영화 j의 선호도 예측
Fig. 1 Preference prediction of newly released movie j based on words

본 연구에서는 영화를 대상 품목으로 사용하였으며 제안한 선호 단어 기반의 시스템은 InfoFinder [5], LIBRA [6]와 유사하게 영화 문서를 분석하여 의미 있는 단어를 영화의 속성으로 추출하고 추출된 속성을 기반으로 새로운 영화에 대한 사용자의 선호도를 예측한다. 다만 기존의 InfoFinder나 LIBRA는 사용자가 대상 품목을 선택할 것인지 그렇지 않을 것인지 만을 판단할 뿐 선호도를 수치화하지 않았다. InfoFinder는 사용자의 선호 여부를 결정하는 의사 결정 트리를 생성하고, 이를 검색 질의로 변환하여 이에 해당하는 품목만을 추천하였으며 LIBRA 또한 훈련 항목에는 1부터 10까지의 평가 값이 할당되어 있지만 대상 품목에 대해 긍정적인가 부정적인가 만을 판단한다. 본 연구에서는 사용자가 대상 품목에 대해 보일 선호도를 구체적으로 수치화한다.

선호 단어 기반의 시스템은 영화 문서를 구성하는 단어를 영화 속성으로 정의하였다. 영화 문서를 구성하는 각 단어에 대해 사용자 선호도를 반영하는 상관 선호도를 계산하여 이를 기준으로 사용자의 영화에 대한 평가 정보를 예측할 수 있는 단어 집합을 추출한다. 이를 사용자의 선호도를 대표하는 "선호 단어(Preference Word) 집합"이라 한다. 사용자의 프로파일은 선호 단어 집합으로 구성되며 대상 사용자에 대한 새 영화에 대한 예측 평가 값은 사용자의 선호 단어 집합과 영화 문서를 구성하는 단어 간의 연관성 분석에 의해 결정한다.

3.2 문서 유사도 기반 시스템 (DOCSIM)

본 논문에서 제안하는 선호 단어 기반의 선호도 예측 시스템과의 성능 비교를 위해 영화 문서를 활용하는 문서 유사도 기법을 구현하였다. 이는 문서 검색 분야에서 널리 활용되는 방법으로 문서 유사도의 계산은 문서에 포함된 단어들의 문서-단어 가중치에 기반을 둔 방법을 사용하였다 [3]. 다음은 두 영화 i, j 간의 문서 유사도 ($DocSim_{i,j}$)이다.

$$DocSim_{i,j} = \frac{\vec{i} \cdot \vec{j}}{|\vec{i}| \times |\vec{j}|} \tag{1}$$

식 (1)로 계산된 영화 문서간의 유사도를 기반으로 사용자 a의 영화 i에 대한 예측 선호도는 CF 시스템에서 사용되는 식을 변형하였다 (식 2).

$$\rho_{a,j} = \bar{r} + \frac{\sum_{i=1}^{n}(r_{a,i} - \bar{r}_a) \times DocSim_{i,j}}{\sum_{i=1}^{n} DocSim_{i,j}} \tag{2}$$

$r_{a,i}$는 사용자 a가 관람한 영화 i에 정한 평가이며, $\bar{r}_a$는 사용자 a가 과거에 관람한 모든 영화에 대한 평가치 가중 평균이다.

3.3 단순 베이지안 분류기(Naïve Bayesian Classifier: NB)

새 영화에 대한 평가 값을 예측하기 위해 자주 사용하는 방법으로 단순 베이지안 분류기 [19]가 있다. 이 방법은 기존에 사용자가 관람한 영화를 구성하는 단어와 영화의 평가 값과의 확률 관계에 기반을 둔다. 새로운 영화가 주어졌을 때, 새 영화에 대한 평가치는 단순 베이지안 분류기에 의해 다음과 같이 계산한다.

$$V_{NB} = argmax_{j=1}^{c} P(v_j) \prod_{k=1}^{m} P(w_k \mid v_j) \tag{3}$$

위 식에서 c는 영화에 대한 평가 분류로 주로 선호와 비선호의 2개 등급으로 표시하지만, 본 연구에서는 보다 정확한 평가 값을 예측하기 위해 영화의 평가 값인 0부터 5까지의 6개 등급을 사용하였다. m은 새 영화를 구성하는 단어의 수이며 $P(v_j), P(w_k \mid v_j)$는 다음과 같이 정의된다.

$$P(v_j) = \frac{|Movie_j|}{|Trains|}$$

$$P(w_k \mid v_j) = \frac{n_k + 1}{n + |Vocabulary|}$$

$Movie_j$는 훈련 집합에 속한 영화 중에서 평가 값이 j인 영화 집합이며, $Trains$는 사용자가 과거에 관람한 영화로 구성되는 훈련 집합이다. $Vocabulary$는 훈련 집합에 속한 영화를 구성하는 단어들의 집합이다. n은 집합 $Text_j$에 속한 단어의 수이며, $Text_j$ 는 평가 값이 j인 문서를 하나로 합한 문서이다. n_k는 단어 w_k의 $Text_j$에서의 단어 빈도수를 의미한다.

3.4 선호 단어 집합 기반 시스템 (Preference word-based system: PWBS)

영화 문서에 포함된 단어와 영화 평가 정보와의 상관관계 분석을 위해 영화 문서 단어의 수치화가 필요하다. 본 연구에서는 단어의 수치 값으로 단어의 문서 내에서의 중요도를 의미하는 문서-단어 가중치를 사용한다. 문서-단어 가중치는 0과 1사이의 값을 갖도록 정규화하며 사용자의 영화에 대한 평가 점수 또한 0.0에서 1.0 사이의 값으로 정규화 한다. 정

규화 된 문서-단어 가중치와 평가 점수를 이용하여 단어 t와 평가 정보(r_a)간의 상관계수(C_{t,r_a})를 계산한다 ($-1.0 \le C_{t,r_a} \le 1.0$). 선호 단어는 단어의 문서-단어 가중치와 평가 정보 간의 상관관계에 따라 결정되며, 상관 계수의 절대 값의 크기를 통해 단어의 중요도를 가늠할 수 있다.

$$C_{t,r_a} = \frac{\sum_{d=1}^{n} w_{d,t} \cdot r_{a,d} - \sum_{d=1}^{n} w_{d,t} \cdot \sum_{d=1}^{n} r_{a,d}}{\sqrt{\left(\sum_{d=1}^{n} w_{d,t}^2 - \left(\sum_{d=1}^{n} w_{d,t}\right)^2\right) \cdot \left(\sum_{d=1}^{n} r_{d,t}^2 - \left(\sum_{d=1}^{n} r_{d,t}\right)^2\right)}} \quad (4)$$

n은 단어 t를 포함하고 있는 영화 문서의 개수이며, $r_{a,d}$는 사용자 a가 영화 d에 부여한 평가 점수이다. $w_{d,t}$는 단어 t의 영화 문서 d에서의 문서-단어 가중치이다.

그림 2는 사용자 a의 영화 j에 대한 선호도 계산을 위한 시스템의 전체 구조이다. 사용자가 과거에 본 영화에 담긴 단어들은 모두 고유의 선호 상관 계수(C_{t,r_a})값을 할당 받게 된다. 할당된 선호 상관 계수의 절댓값($|C_{t,r_a}|$)이 크다는 것은 해당 단어가 사용자의 평가 정보와의 연관성이 크다는 것을 의미한다. 새 영화 j에 포함되어 있는 단어와 사용자의 선호 단어 목록에 공통으로 속한 단어들의 "상대 단어 선호도" ($\tau_{a,t}$)를 계산한다 (식 5). 상대 단어 선호도는 해당 단어를 포함한 새 영화에 대해 사용자의 평균 평가 점수를 기준으로 선호도의 값이 평균보다 높을지 낮을지를 결정한다. 새 영화 j에 대한 최종 예측 선호도($\rho_{a,j}$)는 대상 영화 문서 j와 사용자 a의 선호 단어 목록에 공통으로 속한 단어들의 상대 단어 선호도($\tau_{a,t}$)로 결정한다 (식 6).

$$\tau_{a,t} = \frac{(w_{j,t} - \bar{w}_{d,t})}{\sigma_{w_{d,t}}} \times \sigma_{r_a} \quad (5)$$

$$\rho_{a,j} = \bar{r}_a + \frac{\sum_{t \in T_a, C_{t,r_a} \ge C} \tau_{a,t} \times C_{t,r_a}}{\sum_{t \in T_a, C_{t,r_a} \ge C} C_{t,r_a}} \quad (6)$$

각 항목은 다음과 같이 계산된다.

$$\sigma_{w_{d,t}} = \sqrt{\frac{\sum_{d=1}^{m} (w_{d,t} - \bar{w}_{d,t})^2}{m-1}}, \qquad \bar{w}_{d,t} = \frac{\sum_{d=1}^{m} w_{d,t}}{m}$$

$$\sigma_{r_a} = \sqrt{\frac{\sum_{d=1}^{m} (r_{a,d} - \bar{r}_a)^2}{m-1}}, \qquad \bar{r}_a = \frac{\sum_{d=1}^{m} r_{a,d}}{m}$$

영화 d는 사용자 a가 이미 평가한 영화의 훈련 집합(D_a)에 속한 문서이다. $w_{d,t}$는 영화 문서 d에 포함된 단어 t의 문서-단어 가중치이며, $\bar{w}_{d,t}$는 $w_{d,t}$의 평균이며, $\sigma_{w_{d,t}}$는 $w_{d,t}$의 표준편차이다. $w_{j,t}$ 는 새 영화 j에 포함된 단어 t의 문서-단어 가중치이다 ($d \in D_a$, $j \notin D_a$). $r_{a,d}$는 사용자 a의 영화 d에 대한 평가 값이며, $\bar{r}_{a,d}$는 $r_{a,d}$의 평균이며, $\sigma_{r_{a,d}}$는 $r_{a,d}$의 표준 편차이다. m은 사용자가 관람한 영화 중에서 단어 t를 포함하는 영화의 편수이다. 단어 선호 상관계수 (C_{t,r_a})는 문서-단어 가중치($w_{d,t}$)와 사용자 a가 관람한 영화의 평가 정보(r_a)와의 상관 계수이다. 집합 T_a는 사용자 a의 선호 단어 목록 집합으로 그 단어 중에서 한계값이 일정(C) 이상인 단어만을 사용한다.

```
// 사용자 a가 관람한 영화(D_a)에 속한 단어들의 선호상관계수를 계산.
// T_a 는 사용자 a가 평가한 영화 문서에 포함된 단어들의 집합.
// r_a 는 사용자 a가 단어 t를 포함한 영화에 대해 부여한 평가 값
for all t ∈ T_a
{
    C_{t,r_a} = Correlation_Coefficient(t, r_a)
}
sum = 0;
sum_corr = 0;
for all t ∈ T_a
{
    // C는 한계값.
    if ( C_{t,r_a} > C )
    {
        // 상대 단어 선호도를 계산.
        τ_{a,t} = (w_{j,t} − w̄_{d,t}) / σ_{w_{d,t}} × σ_{r_a} ;
        sum = sum + τ_{a,t} × C_{t,r_a} ;
        sum_corr = sum_corr + C_{t,r_a};
    }
}
// 최종 예측 값은 가중치 평균으로 결정.
ρ_{a,j} = r̄_a + sum / sum_corr
return ρ_{a,j};
```

그림 2. 사용자 a의 새 영화 j에 대한 선호도 예측 시스템
Fig. 2. Prediction of user's preference for a new movie

3.5 비매칭 문제 해결을 위한 확장 문서 유사도 시스템 (EXDOCSIM)

문서 유사도는 두 문서 간에 공통적으로 존재하는 단어간의 문서-단어 가중치를 사용하여 계산하는데 그림 3의 두 문서 i, j에는 공통적으로 포함되는 단어가 많다 (회색과 빗금친 부분은 문서 i, j에 단어 k가 포함된 경우를 의미한다). 공통으로 들어가는 단어가 많은 경우에는 문서의 유사도 수치를 신뢰할 수 있다. 다른 경우로 그림 4의 문서 p와 문서 q는 공통 단어가 적기 때문에 유사도에 의미를 부여하기 힘들다. 심할 경우는 두 문서간의 공유 단어가 전혀 없을 수도 있다.

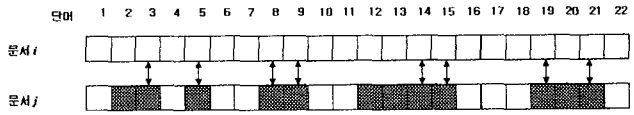

그림 3. 두 문서간의 공통 단어가 많은 경우
Fig. 3. The case that there are many common attributes between two documents.

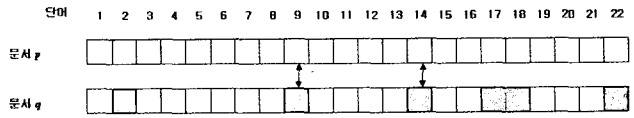

그림 4. 두 문서간의 공통 단어가 적은 경우
Fig. 4. The case that there are few common attributes between two documents.

그림 4에서 문서 p의 벡터 $\vec{p}$는 $0, 0, w_{p,3}, 0, w_{p,5}, 0, 0, w_{p,8}$,

$w_{p,9}, w_{p,10}, 0, 0, 0, w_{p,14}, w_{p,15}, 0, 0, 0, w_{p,19}, 0, w_{p,21}, 0$ 로 표현되며, 문서 q의 벡터 $\vec{q}$ 는 $0, w_{q,2}, 0, 0, 0, 0, 0, 0, w_{q,9}, 0, 0, 0, 0, w_{q,14}, 0, 0, w_{q,17}, w_{q,18}, 0, 0, 0, w_{q,22}$ 로 표현된다. 벡터로 표현된 두 문서의 유사도 $DocSim_{p,q}$는 서로 공유한 요소를 사용하여 내적으로 표현한다. 즉,

$$\vec{p} \cdot \vec{q} = \sum_{k} w_{p,k} \times w_{q,k}$$

이 된다. 이 값을 문서의 유사도로 사용할 수 있겠지만, 유사도의 신뢰도를 높일 필요가 있다.

본 연구에서는 기호 네트워크를 통하여 두 문서간의 공통 단어의 희소성 문제를 보강한다. 문서 p, q에 포함된 단어를 확장함으로써 공통 단어의 수를 늘리는 것이다. 기호 네트워크를 통하여 문서 p에는 포함되지 않던 단어 2, 17, 22를, 문서 q에서는 단어 3, 8, 15, 19가 확장되었다 (그림 5).

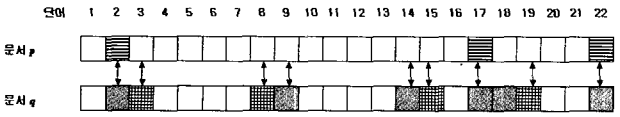

그림 5. 기호 네트워크를 이용하여 확장된 문서 유사도
Fig. 5 Expanded document similarity using semiotic network.

기호 네트워크에 의해 확장된 단어를 "가상 단어"라고 하며, 이 가상 단어 집합을 이용하여 얻는 유사도의 값을 "가상 유사도(virtual similarity)"라 한다. 최종적인 확장 유사도는 기본 유사도($DocSim_{p,q}$)와 가상 유사도($VDocSim_{p,q}$)의 가중치 합으로 정한다.

$$VDocSim_{p,q} = ew_{p,2} \cdot w_{q,2} + w_{p,3} \cdot ew_{q,3} + w_{p,8} \cdot ew_{q,8} + ew_{p,15} \cdot ew_{q,15} + ew_{p,17} \cdot w_{q,17} + w_{p,19} \cdot ew_{q,19} + ew_{p,22} \cdot w_{q,22}$$

이를 일반화하면 두 문서 p, q 간의 가상 유사도($VDocSim_{p,q}$)는 다음과 같다.

$$VDocSim_{p,q} = \sum_{l=1}^{s} (w_{p,l} \times ew_{q,l}) + \sum_{m=1}^{t} (ew_{p,m} \times w_{q,m})$$

$ew_{q,l}$는 문서 p에는 포함되었지만 문서 q에는 포함되지 않은 단어 l의 문서 q에 대한 추정 문서 단어 가중치 (estimated document-term weight value: ETF-IDF)이다 ($l \in T_p$, $l \notin T_q$). 같은 방법으로 $ew_{p,m}$는 문서 q에는 소속되었지만, 문서 p에는 소속되어 있지 않는 단어 m의 문서 p에 대한 추정 문서 단어 가중치이다 ($m \in T_q$, $m \notin T_p$). 추정 문서-단어 가중치는 문서 내에 포함되어 있지 않지만, 기호 네트워크 분석을 통해 문서에 포함된 다른 단어와의 유사도를 기반으로 추정한 값이다. 문서 q의 가상 단어 l에 대한 추정 문서-단어 가중치($ew_{q,l}$)은 기호 네트워크에서 단어 l과 연관성이 높고 문서 q에 포함된 단어 w의 문서-단어 가중치의 연관가중치 평균값으로 정한다 ($l \notin T_q$, $w \in T_q$). 단어 w 중에서 문서 p에 속하는 단어는 제외하는데, 이는 기본 유사도 계산에서 이미 반영되기 때문이다 ($w \notin T_p$). 물론 $ew_{p,m}$ 경우도 동일한 방법으로 계산한다 ($m \notin T_p$, $w \in T_p$, $w \notin T_q$) (식 7).

$$ew_{q,l} = \frac{\sum_{w=1}^{\Omega} (w_{q,w} \times R_{l,w})}{R_{l,w}} \quad , \quad ew_{p,m} = \frac{\sum_{w=1}^{\Omega} (w_{p,w} \times R_{m,w})}{R_{m,w}} \tag{7}$$

$R_{l,w}$, $R_{m,w}$는 각각 단어 l과 m에 대해서 기호 네트워크에서 서로 연결되어 있는 단어 w에 대한 유사도이다. 단어 간의 유사도의 크기는 Jaccard method($\zeta(l,w)$)를 사용하였으며 두 단어간의 유사도가 일정 값(C) 이상인 경우에 유사 단어로 간주하였다 [24].

$$R_{l,w} = \begin{cases} \zeta(l,w) & \text{if } \zeta(l,w) > 0 \\ 0 & otherwise \end{cases},$$

$$\zeta(l,w) = \frac{P(l,w)}{P(l) + P(w) - P(l,w)}$$

$P(l)$, $P(w)$는 각각 단어 l과 단어 w의 MLE 값이며, $P(l,w)$는 두 단어가 공통으로 포함되는 경우의 MLE이다. 확장 단어를 기반으로 한 두 영화 문서 간의 확장 유사도 ($ExDocSim_{p,q}$)는 문서 유사도($DocSim_{p,q}$)와 가상 유사도 ($VDocSim_{p,q}$)의 가중치 합으로 결정된다. 가중치 α, β의 값은 경험적으로 적절히 결정한다 (식 8).

$$ExDocSim_{p,q} = \alpha \cdot DocSim_{p,q} + \beta \cdot VDocSim_{p,q} \tag{8}$$

식 (9)는 두 문서간의 확장 유사도를 적용하여 영화에 대한 선호도를 예측하는 방법으로 식 (2)의 $DocSim_{p,q}$ 대신에 $ExDocSim_{p,q}$을 사용한다.

$$\rho_{a,j} = \bar{r}_a + \frac{\sum_{i=1}^{n} (r_{a,i} - \bar{r}_a) \times ExDocSim_{i,j}}{\sum_{i=1}^{n} ExDocSim_{i,j}} \tag{9}$$

4. 실험 결과

4.1 평가 방법

본 연구에서는 예측 시스템의 성능 측정을 위해 예측의 정밀도를 나타내는 평균절대오차(MAE), 분류의 정밀도를 나타내는 F1 척도와 ROC (Receiver Operating Characteristics) 감도를 사용하였다 [25].

F1 척도는 재현율(recall)과 정확율(precision)이 서로 상대적인 의미로 어느 한쪽의 값이 올라가면 다른 값은 내려가므로 어느 한쪽의 값으로 시스템의 성능을 결정할 수 없으므로 재현율과 정확율을 모두 고려하여 성능을 결정하는 방법이다. 그 값은 $F1 = \frac{2 \times \text{재현율} \times \text{정확율}}{\text{재현율} + \text{정확율}}$로 계산된다.

ROC 감도 또한 재현율과 정확율 방법의 대안으로 제안되었으며 예측 시스템에서는 각 품목에 사용자의 예측 평가치를 계산한 후 한계값을 설정하여 한계값 이상의 품목은 사용자의 선호 품목으로 그렇지 않은 품목은 비선호 품목으로 예측하여 정확도를 계산한다. 본 실험에서 사용자는 영화에 대

해 0부터 5까지의 평가 값을 부여하였으며 분류 정밀도 측정을 위하여 평가치 중에서 0부터 3까지는 관심 없는 영화로, 4, 5를 받은 경우에만 관심 있는 영화로 분류하였다. 이 경우를 ROC-4 감도법이라고 한다. 만약 0, 1, 2를 무관심으로, 3, 4, 5를 관심으로 하는 경우는 ROC-3 감도라 한다. F1-3 척도와 F1-4 척도도 이와 유사하다. ROC 감도의 값의 범위는 0.0부터 1.0까지로, 0.5는 무작위 추천의 경우이며 1인 경우는 완벽한 예측 시스템을 말한다 [8].

4.2 실험 데이터

본 연구에서 사용하는 데이터는 DEC(Digital Equipment Corporation)의 시스템 연구센터(Systems Research Center: SRC)에서 사용한 연구 목적의 데이터베이스로 1995년부터 1997년까지 18개월 동안 수집되었으며 CF 기반의 영화 추천 시스템 구축에 사용되었다[1]. EachMovie 데이터는 사용자의 영화 평가 정보와 더불어 영화의 장르, 영화 홈페이지 URL, IMDb사의 영화 URL 등을 포함하고 있으며, 본 논문에서는 EachMovie 데이터의 사용자 평가 정보와 더불어 해당 영화에 IMDb[2]사의 영화 홍보 사이트의 내용을 영화의 속성 정보로 활용하였다.

EachMovie의 데이터에서 20편 미만의 영화를 관람한 사용자와 IMDb 사에서 영화 정보를 얻을 수 없는 영화는 제외하였다. 새로운 영화에 대한 사용자 선호도 예측 시스템을 시뮬레이션 하기 위해 전체 1,620편의 영화에서 1편을 새로운 영화로 선정하고 나머지 영화는 사용자의 프로파일 생성을 위한 훈련 영화로 사용하였다. 이러한 방식은 교차검증(Cross Validation) 기법의 극단적인 형태로서 전체 N편의 영화 중에서 1편만을 테스트 집합으로 지정하고 나머지 $N-1$ 개의 데이터를 훈련 집합으로 활용한다 [26]. 테스트 실행 회수는 모두 N번으로 i번째 테스트에서는 영화 i를 제외한 모든 영화가 훈련 집합으로 사용된다. EachMovie 데이터를 사용한 Basilico [10] 등이 전체 영화 중에서 일부를 샘플링하여 사용한 반면에 본 연구에서는 전 영화를 대상으로 하였다.

4.3 시스템의 예측 성능 비교

본 연구에서 제안하는 사용자 선호 단어 기반의 예측 기법, 확장 문서 유사도 기법과 기존의 문서 유사도 기반 기법, 단순 베이지안 분류기를 서로 비교하였다 (표 1). 평균절대오차(MAE)는 그 값이 작을수록 좋은 시스템이며, F1 척도와 ROC 감도는 값이 높을수록 좋은 성능을 의미한다. 표에서 굵게 표기한 것은 가장 좋은 것을 의미한다.

표 1 각 시스템 성능 비교
Table 1. Performance comparison of different methods

기법	MAE	ROC-3	ROC-4	F1-3	F1-4
DOCSIM	1.110307	0.696655	**0.676734**	0.802177	0.582673
EXDOCSIM	1.094520	0.703045	0.6740560	**0.822690**	**0.629089**
PWBS	**1.086530**	**0.704368**	0.667817	0.807569	0.512276
NB	1.177480	0.668442	0.655989	0.821550	0.610727

선호 단어 시스템(PWBS)과 확장 문서 유사도 시스템(EXDOCSIM)이 전체적으로 기존 문서 유사도 시스템(DOCSIM)과 단순 베이지안 분류기(NB)에 비하여 성능 면에서 상대적으로 좋거나 비슷한 성능을 보였다.

4.4 추천 설명

문서 유사도에 의한 선호도 예측은 추천 영화와 사용자가 과거에 관람한 추천 영화와의 유사도를 고려하므로 사용자에게 왜 이 영화를 추천하는가에 대한 설명이 부족하다. 물론 영화 대 영화의 관점에서 설명할 수 있지만, 영화의 어떤 점이 영향을 미쳤는가는 설명하기 어렵다. Mooney 등은 LIBRA에서 선호 강도가 높은 단어들을 포함한 도서를 리스트의 상위 순서에 배치시키고 사용자에게 책의 추천 이유를 사용자들의 강도가 높은 단어가 포함된 도서의 목록을 제시하였다 [6]. 본 연구에서는 단어의 선호 상관도와 단어의 중요도를 활용하여 예측의 근거를 설명한다. 사용자 프로파일이 구축되면 시스템은 선호 단어 집합을 기반으로 새로운 영화에 대한 평가 값을 예측한다.

표 2는 어떤 사용자의 선호 단어 리스트를 보여주고 있다. "영화 수"는 단어의 선호 상관 계수를 결정하기 위해 사용된 영화의 수를 의미하며, "선호 상관 계수"는 대상 영화에 포함된 단어의 중요도 즉, 문서-단어 가중치와 영화 평가 점수와의 상관관계 정도를 의미한다. 상관 선호 값이 1에 가까운 것은 해당 단어가 중요도가 높을수록 평가 값이 높을 것임을 의미하며 이러한 단어는 "선호단어"로 분류된다. 선호 값이 -1이 가깝고 단어 중요도가 높으면 평가 값은 낮게 예측되는데 이러한 단어는 "혐오 단어"로 분류된다 (표 3).

표 2. 어떤 사용자의 새 영화 j에 대한 선호도 계산을 위한 선호 단어와 문서 단어 가중치
Table 2. The preference correlation coefficient and TF-IDF values of words to determine the user preference for a new movie j.

선호 단어	선호 상관 계수	새 영화 j의 선호단어의 문서-단어 가중치
ANIMATION	0.547701	0.242411
APPRENTICE	-0.450898	0.014355
BLOCKBUSTER	-0.420438	0.011093
BOB	-0.353978	0.022511
PETER	-0.431822	0.010963
TEAM	-0.666319	0.341655
VOICE	0.273943	0.212892

표 3. 상관 선호 계수와 단어 중요도 기반의 선호/혐오 단어 결정
Table 3. Determining a like/dislike word based on the correlation coefficient and TF-IDF.

	선호 상관 계수 값 (1에 가까움)	선호 상관 계수 값 (-1에 가까움)
중요도 높음	선호 단어	혐오 단어

선호 상관 계수의 값이 절대값이 0에 가까운 것은 해당 단어가 영화 평가에 대한 관련성이 낮음을 의미한다. 단어 "ANIMATION", "VOICE"는 선호 상관 계수가 높으면서 단어의 중요도인 문서-단어 가중치 값이 높은 선호 성향 단어이

1) http://www.research.compaq.com/SRC/eachmovie/#obtaining The EachMovie Dataset remained available until October 2004 when it was finally retired

2) http://www.imdb.com

며, "TEAM"은 문서-단어 가중치가 높지만, 선호 상관 계수의 값이 음의 상관 계수이므로 혐오 성향 단어임을 보여 준다.

4.5 속성 비매칭 문제의 개선

그림 6에서는 두 영화간의 속성 비매칭 문제로 인해 영화간의 유사도를 계산하지 못하여 영화에 대한 선호도 예측이 불가능한 사용자의 비율 변동 추이를 보여주고 있다. 확장 문서 유사도를 적용하는 경우가 기존 문서 유사도 기법에 비해 추천 불가 사용자의 비율을 많이 줄여주고 있음을 볼 수 있다.

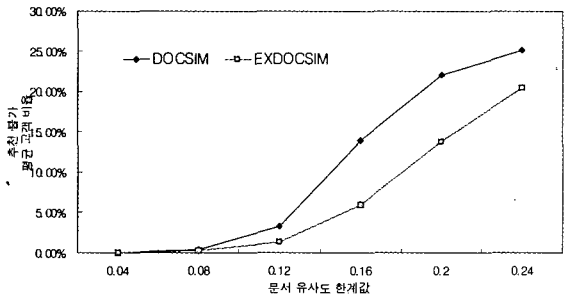

그림 6. 비매칭 문제로 인한 추천 불가 사용자 비율의 변화
Fig. 6. Ratio change of users impossible to be recommended for a new movie due to the mismatching problem.

그림 7은 확장 문서 유사도를 적용한 시스템과 그렇지 않은 시스템의 성능 변화를 보여주고 있다. 숫자 5, 10은 두 문서간에 공통으로 속한 단어의 수가 각각 5개, 10개 이하인 경우에만 가상 문서 유사도를 적용하는 것을, All은 항상 가상 문서 유사도를 적용하는 경우이다. 공통 단어 수가 5 개 이하인 경우(USER-EXDOCSIM-5)에서 절대 평균 오차의

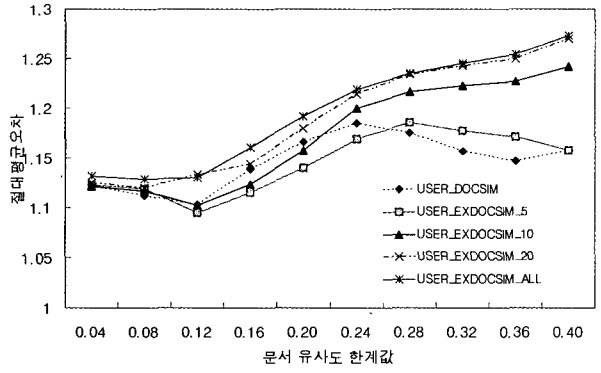

그림 7. 확장 문서 유사도를 적용한 시스템의 성능 변화
Fig. 7. Performance of expanded document similarity methods.

성능이 개선되었으나 공통 단어의 수의 한계값을 늘린 시스템(USER-EXDOCSIM-10, USER-EXDOCSIM-20, USER-EXDOCSIM-ALL)은 절대 평균 오차의 성능이 더 나빠졌다. 이는 가상 유사도의 값의 지나친 확장은 순수 문서 유사도에 의한 선호도 예측을 방해하는 것을 알 수 있다. 이를 the drift problem이라고 하며 과거에도 관찰된 현상이다 [27].

5. 결 론

본 논문에서는 사용자가 과거에 관람한 영화의 홍보 문서를 분석하여 사용자의 선호도를 반영한 선호 단어를 추출하고 이 선호 단어를 이용하여 새로이 개봉되는 영화에 대한 사용자의 평가를 예측하는 시스템을 제안하였다. 제안된 시스템은 기존 문서 유사도 기반의 시스템과 단순 베이지안 분류기에 비해 개선된 성능을 보여주었다. 이 시스템은 평가값을 예측할 뿐만 아니라, 사용된 선호 단어 집합과 그 상관 선호도 값을 제시함으로써 시스템의 평가 정보 예측에 대한 근거를 설명하였다. 또한 영화의 유사성 계산 방법에 사용되는 문서 유사도 계산에서 발생하는 속성 비매칭 문제를 해결하기 위해 기호 네트워크에 기반을 둔 확장 문서 유사도를 사용하였다. 이로써 기존 품목들과 공통 속성이 적은 품목이라 할지라도 그 유사도를 추정할 수 있다.

선호 단어 기반 예측 시스템에서 선호도 계산에 사용된 영화의 수와 단어의 선호 상관 계수의 값이 모두 중요하다. 앞으로의 연구에서는 시스템의 예측 정확도를 향상시키기 위해 선호 단어의 두 성분을 적절히 혼용하는 방법을 찾을 것이다. 또한 선호 단어 시스템에 기호 네트워크를 적용하기 위한 모델을 고안할 것이다.

참 고 문 헌

[1] P. Resnick and H. R. Varian, "Recommender systems, Communications of the ACM", Vol.40, No.3, pp. 56-58, 1997.

[2] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry", Communications of the ACM, Vol.35, No. 12, pp. 61-70, 1992.

[3] J. Zobel and A. Moffat, "Exploring the Similarity Space", SIGIR Forum, Vol. 32, No. 1, pp. 18-34, 1998.

[4] K. Lang, NewsWeeder: learning to filter netnews, Proceedings of the 12th International Conference on Machine Learning, pp. 331-339, 1995.

[5] B. Krulwich and C. Burkey, "The InfoFinder agent: learning user interests through heuristic phrase extraction", IEEE Intelligent systems, Vol. 12, No. 5, pp. 22-27, 1997.

[6] R. J. Mooney and L. Roy, Content-based book recommending using learning for text categorization, Proceedings of DL-00, 5th ACM Conference on Digital Libraries, pp. 195-204, 2000.

[7] M. J. Pazzani and D. Billsus, "Learning and Revising User Profiles: The Identification of Interesting Web Sites", Machine Learning, Vol. 27, No. 3, pp. 313-331, 1997.

[8] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, An algorithm framework for performing collaborative filtering, Proceedings of the Conference of Research and Developments in Information Retrieval, pp. 219-233, 1999.

[9] E. Terra and C. L. A. Clarke, Scoring missing terms in information retrieval tasks, CIKM '04: Proceedings of the Thirteenth ACM conference on Information and knowledge management, pp.50-58, 2004.

[10] J. Basilico and T. Hofmann, Unifying collaborative and content-based filtering, Twenty-first international conference on Machine learning, 2004.
[11] P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm and J. Riedl, GroupLens: An Open Architecture for Collaborative Filtering of Netnews, Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, pp. 175-186, 1994.
[12] H. L. Stead, M. Rosenstein, and G. Furnas, Recommending and Evaluating Choices in A Virtual Community of Use, Proceedings of the CHI-95 Conference on Human Factors in Computing systems, pp.194-201, 1995.
[13] U. Shardanand and P. Maes, Social information filtering: algorithms for automating word of mouth, Proceedings of CHI'95 Conference on Human Factors in Computing systems, pp. 210-217, 1995.
[14] D. Greening, Building Consumer Trust with Accurate Product Recommendations, LikeMinds White Paper LMWSWP-210-6966, 1997.
[15] J. L. Herlocker, J. A. Konstan, and John Riedl, Explaining collaborative filtering recommendations, Computer Supported Cooperative Work, pp. 241-250, 2000.
[16] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, GroupLens: Applying Collaborative Filtering to Usenet News, Communications of the ACM, Vol. 40, No. 3, pp. 77-87, 1997.
[17] G. Salton, "Associative document retrieval techniques using bibliographic information", Journal of the American Society for Information Science, Vol. 10, No. 4, pp. 440-457, 1963.
[18] G. Fischer and C. Stevens, Information access in complex, poorly structured information spaces, Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 63-70, 1991.
[19] Mitchell T., Machine Learning, McGraw-Hill, New York, 1997.
[20] M. Balabanovic and Y. Shoham, "Combining Content-Based and Collaborative Recommendation", Communications of the ACM, Vol. 40, No. 3, 1997.
[21] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database", Journal of Lexicography, Vol. 3, No. 4, pp. 234-244, 1990.
[22] S. Y. Lee, S. S. Choi and B. R. Moon, Search Improvement by Genetic Algorithms with a Semiotic Network, GECCO 2002: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 1126-1132, 2002.
[23] J. Xu and W. B. Croft, "Improving the effectiveness of information retrieval with local context analysis", ACM Transactions on Information systems, Vol.18, No. 1, pp. 79-112, 2000.
[24] P. N. Tan, V. Kumar, and J. Srivastava, Selecting the right interestingness measure for association patterns, KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 32-41, 2002.
[25] J. L. Herlocker, J. A. Konstan, L. G. Terveen and J. T. Riedl, "Evaluating collaborative filtering recommender systems", ACM Transactions on Information systems, Vol. 22, No. 1, pp. 5-53, 2004.
[26] R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, pp. 1137-1145, 1995.
[27] C. J. Crouch, D. B. Crouch, Q. Chen, and S. J. Holtz, "Improving the retrieval effectiveness of very short queries", Information Processing and Management, Vol. 38, No. 1, pp. 1-36, 2002.

저 자 소 개

최윤석(Yoon-Seok Choi)
1996년 : 중앙대학교 전자계산학(학사)
1998년 : 중앙대학교 컴퓨터공학(석사)
1999년~현재 : 서울대학교 컴퓨터공학부 박사과정

관심분야 : CRM, 데이터마이닝, 기계학습, 유전알고리즘.
E-mail : yschoi@soar.snu.ac.kr

문병로(Byung-Ro Moon)
1985년 : 서울대학교 계산통계학(학사)
1987년 : 한국과학기술원 전산학(석사)
1994년 : Pennsylvania주립대 전산학 (박사)
1997년~현재 : 서울대학교 컴퓨터공학부 부교수

관심분야 : 최적화, 알고리즘설계, 유전알고리즘, 진화연산, 복잡계
E-mail : moon@soar.snu.ac.kr