JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 10, pp. 163-170, October 2023
https://doi.org/10.9708/jksci.2023.28.10.163

Jaccard Index Reflecting Time-Context for User-based Collaborative Filtering

Soojung Lee*

*Professor, Dept. of Computer Education, Gyeongin National University of Education, Anyang, Korea

[Abstract]

The user-based collaborative filtering technique, one of the implementation methods of the recommendation system, recommends the preferred items of neighboring users based on the calculations of neighboring users with similar rating histories. However, it fundamentally has a data scarcity problem in which the quality of recommendations is significantly reduced when there is little common rating history. To solve this problem, many existing studies have proposed various methods of combining Jaccard index with a similarity measure. In this study, we introduce a time-aware concept to Jaccard index and propose a method of weighting common items with different weights depending on the rating time. As a result of conducting experiments using various performance metrics and time intervals, it is confirmed that the proposed method showed the best performance compared to the original Jaccard index at most metrics, and that the optimal time interval differs depending on the type of performance metric.

▸ **Key words**: Jaccard Index, Data Sparsity Problem, Similarity Measure, Collaborative Filtering, Time-aware Recommender System

[요 약]

추천 시스템의 구현 방식들 중 하나인 사용자 기반의 협력 필터링 기법은 유사한 평가 이력을 가진 이웃 사용자들의 산출을 기반으로 하여, 이들의 선호 항목들을 추천한다. 그러나 공통된 평가 이력이 적을 경우에 추천의 질이 현저히 저하되는 데이터 희소성 문제를 근본적으로 갖고 있다. 이러한 문제의 해결을 위하여 많은 기존 연구에서 자카드 계수를 유사도 척도와 접목하는 다양한 방법들을 제안해 왔다. 본 연구에서는 자카드 계수에 시간 인지 개념을 도입하여 공통 항목의 평가 시간에 따라 다른 비중으로 가중합하는 방안을 제시한다. 다양한 성능 척도와 시간 주기를 활용하여 실험을 수행한 결과, 제안 방법이 대부분의 척도에서 원래의 자카드 계수에 비해 가장 우수한 성능을 보였으며, 최적의 시간 주기는 성능 척도의 종류에 따라 다름을 확인하였다.

▸ **주제어**: 자카드 계수, 데이터 희소성 문제, 유사도 척도, 협력 필터링, 시간 인지 추천 시스템

• First Author: Soojung Lee, Corresponding Author: Soojung Lee
*Soojung Lee (sjlee@gin.ac.kr), Dept. of Computer Education, Gyeongin National University of Education
• Received: 2023. 09. 25, Revised: 2023. 10. 16, Accepted: 2023. 10. 17.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

인터넷 활용이 일상적인 현대 사회에서 추천 시스템은 필수 불가결한 요소로서 자리 매김하였다. 다양한 영역의 웹 쇼핑, 엔터테인먼트 사이트 등에서 사용자들의 만족을 극대화하고 정보 획득의 용이함을 실현하기 위하여 추천 알고리즘을 구현 및 활용하고 있다.

이제까지 개발된 여러 추천 알고리즘들 중에서 협력 필터링(collaborative filtering, CF) 방식은 구현이 용이하고 효율적이라는 장점으로 인해 상업 시스템에 가장 널리 활용되어 왔는데, 아마존, 넷플릭스 등이 대표적인 예이다[1][2]. CF는 크게 사용자 기반(user-based)과 항목 기반(item-based) 방식으로 나뉘는데, 그에 따라 인접 이웃(nearest neighbor)의 종류가 달라진다[3]. 전자의 방식은 현 사용자 또는 목표 사용자(target user)와 유사한 행동 양식 또는 선호도를 나타내어 왔던 사용자들을 인접 이웃 집합으로 정하고, 그들이 높은 평가를 부여했던 항목들을 추천한다. 반면에 항목 기반은 현 사용자가 높게 평가하였던 항목들과 유사한 항목들을 인접 이웃들로 간주하고 이들을 추천하는 방식이다. 따라서 CF 시스템에서 유사도 측정은 성능을 결정하는 매우 중요한 이슈이다.

지난 십수년간 CF 관련 연구에서 다양한 유사도 척도들이 개발되었다[3][4]. 전통적으로 사용되어 온 척도들로서 피어슨 상관도, 코사인 유사도, 유클리디안 유사도 등이 대표적이다[5][6]. 이들은 사용자가 항목에 대하여 직접적으로 입력한 평가치를 기반으로 유사도 값을 산출한다. 만약 직접 평가치가 부재할 경우엔, 사용자의 방문 행태, 클릭 회수, 사이트 체류 시간 등을 통해 간접적으로 선호도를 파악하고 이로부터 평가치를 얻을 수 있다[1][3]. 협력 필터링 시스템과 관련되어 유사도 척도들의 장단점은 Fkih 또는 Saranya 외 2인에 의해 상세히 제시되었다[4][5].

CF 시스템의 상태에 적합한 유사도 척도를 선정하고 이의 활용을 통해 최적의 인접 이웃들을 구하는 일은 매우 중요하다. 그러나, 더욱 중요한 것은 과연 해당 시스템이 사용자의 평가치 데이터를 충분히 갖고 있느냐 하는 문제이다. 데이터가 적을 경우 유사도 산출이 불가하거나 산출된 유사도 값을 신뢰하기 어렵기 때문이다. 이러한 데이터 희소성(data sparsity)은 CF 시스템의 고질적인 문제로서 이의 해결을 위해 많은 연구 업적들이 발표되었다[7][8]. 특히, 자카드 계수(Jaccard index)는 두 사용자의 항목 평가 총개수와 대비하여 공통으로 평가한 항목들 개수의 비율을 반영한 수치인데, 이 계수를 기존에 개발된 유사도 척도들과 결합하여 CF 시스템의 데이터 희소성 문제를 해결하기 위한 목적으로 활용한 연구 결과들이 보고되었다[9][10]. 한편 평가데이터, 공통 평가 항목항목 개수 등을 활용하는 것 이외에 CF 시스템의 성능 개선을 위하여 평가 시간 인지(time-aware), 위치 인지(location-aware) 등의 상황 인지(context aware) 기법을 접목하는 연구도 발표되었다[11][12]. 대부분의 평가 시간 인지 CF 시스템은 사용자들의 과거 평가치에 대하여 낮은 가중치를 부여하여 변환한 값을 활용함으로써 과거 평가치의 영향력을 감소시키는 방식을 취한다.

본 연구에서는 기존의 자카드 계수에 시간 인지 개념을 도입한 시간 인지 자카드 계수를 제안한다. 연구 목적은 유사도 측정을 위해 두 사용자의 공통 평가 항목수의 비율을 시간에 따라 다른 비중으로 취급하는 것이 CF 시스템의 성능에 미치는 영향을 파악하기 위함이다. 제안 방법은 주기적으로 사용자 간의 자카드 계수를 산출한 후, 산출 시간에 따라 다른 가중치가 부여된 계수들의 합을 구하고 이를 유사도 값으로 활용한다. 실험을 통하여 시간 주기의 변화에 따른 성능 변화를 측정하고 원래의 자카드 계수 성능과 비교하였으며, 예측 정확도와 추천 정확도 등의 다양한 척도를 활용하여 분석하였다. 분석 결과 제안 방법은 거의 모든 측면에서 가장 우수한 성능을 보였으며, 최적의 시간 주기는 성능 척도의 종류에 따라 달랐다.

논문의 구성은 다음과 같다. 2절에서는 시간 인지 및 자카드 계수 활용의 협력 필터링과 관련된 기존 연구 성과를 소개한다. 3절에서는 제안 방법을 설명하고 4절에서 성능 실험 결과를 제시하며, 5절에서 논문의 결론을 맺는다.

II. Related Works

1. Jaccard Index

사용자 간 또는 항목 간의 유사도를 활용하는 CF 시스템에서 공통 평가 항목수는 유사도 산출의 기반이 되므로 매우 중요한 요소이다. 그러나, 평가 데이터가 충분하지 않을 때 특히 데이터 희소성 문제가 발생할 수 있으므로, 이를 해결하기 위한 많은 연구가 시도되었다.

Default voting은 매우 간단한 방법으로서, 미평가항목에 대하여 기본값을 부여함으로써 밀집된 데이터 환경을 구축할 수 있다[3]. 이밖에 기계 학습(machine learning), 베이지안 다중 대치(Bayesian multiple imputation), 선형 회귀(linear regression) 알고리즘으로 데이터 희소성을 극복하려는 노력을 기울여 왔다[1][3].

사용자들의 미평가 항목에 대한 평가치를 유추하기 위

한 작업은 CF 시스템에서 활용하는 사용자-항목 매트릭스 자체를 보완하려는 것이다. 이는 대개 성능 개선에 도움을 가져오는 것으로 알려졌으나, 대체하는 평가치가 부정확할 수 있으므로 성능 향상에 한계가 있다. 따라서 이러한 방안에서 벗어나 사용자들의 실제 평가치 만을 활용하여 희소 데이터 문제를 해결하려는 노력이 강구되었다.

이러한 노력의 가장 단순한 구현 아이디어는 유사도 산출을 위하여 공통 평가 항목 집합의 크기를 반영, 즉, 집합의 크기와 두 사용자 간 유사도 값이 비례하도록 하는 것으로서 추천 시스템의 성능 개선의 효과가 크다고 알려져 있다[7]. 예를 들어, Herlocker 외 3 인은 산출된 유사도 값에 공통 항목 수에 비례하는 값을 곱하여 최종 유사도를 산출하였다[13]. Kwon 외 3 인에서 언급한 Tanimoto 계수는 전체 평가 항목 수 대비 공통 평가 항목수를 의미한다[14]. 마찬가지로, 자카드 계수 또한 공통 평가 항목수의 비율을 측정하는데[9], 이와 같은 계수들은 사용자의 평가치 자체는 반영하지 않기 때문에 유사도 산출을 위한 보조적인 역할을 한다. 따라서, 희소성 문제를 극복하기 위하여 다양한 기존의 유사도 척도들과 결합하여 최종적인 유사도 값을 산출하는 목적으로 주로 활용되었다[5][7][15].

2. Time-aware Collaborative Filtering

기존의 전통적인 CF 방법들은 대개 사용자의 항목 평가 점수를 활용하여 추천 리스트를 결정하였으나, 성능 향상을 위하여 상황 인지(context-aware) CF 기법이 개발되었다. 이 중에서 시간 인지(time-aware) CF는 평가치가 부여된 시간을 추가적으로 고려하는 방법으로서, 과거 평가치에 대하여 기하급수적으로 감소하는 가중치를 부여하여 해당 평가치의 영향력을 감쇠시키는 것이 전형적이다[12][16].

이와 같은 감쇠율을 적용하는 다양한 연구 결과들이 발표되었는데, H.-Zhen과 Lei의 연구에서는 과거 평가치에 대하여 로지스틱 시간 함수를 적용하여 유사도를 계산하였다[17]. Campos 외 2 인은 시간 인지 추천 시스템에 대하여 전반적으로 사전 연구들을 분석하고, 성능 평가를 위한 다양한 이슈들, 즉, 실험 데이터 분할 방법, 분할 시에 고려할 평가 시간 기반의 데이터 집합 크기 문제, 시간 인지 알고리즘의 성능 평가 및 사용 척도 등에 대하여 언급하였다[11].

평가치에 대하여 시간 가중 함수를 직접적으로 적용하는 방식 외에 클러스터링 기법을 활용하여 시간 인지 기법을 개발한 연구 결과들도 발표되었다. 예로서 항목들을 군집화한 후에 각 군집 별로 다른 감쇠율을 적용한 Ding과 Li의 연구는 초창기의 시간 인지 CF 기법으로서 이후 많은 연구자들의 관심을 유도하였다[18]. 또한 평가 시간을 구간별로 나누고 각 구간별 평가 개수를 고려하여 가중치를 부여한 연구가 He와 Wu에 의해 발표되었다[19].

시간 요소는 문서 태그를 활용한 추천 시스템에서도 고려되어 사용자의 선호 태그를 파악하는데 활용되었는데, 사용자의 선호 문서에 대한 가중치를 조회 시간에 무관하게 운용하는 방법이 제안되었다[20]. 사용자의 리뷰 정보와 선호도 변화에 있어서 시간 함수를 적용하여 분석하고 이를 추천 시스템에 반영하는 연구 결과도 제시되었다[21]. 한편 평가치 자체만을 기반으로 하는 유사도 척도의 성능 단점을 개선하고자 여러 가지 다양한 정보를 통합하는 방식이 연구되어 왔다[22]. 대표적인 정보의 예는 평가 시간의 차이, 평가치의 차이, 평가 순서의 차이, 사용자의 확신도 등이다. 이들 요소 각각의 최적의 비중은 실험을 통해 정하였다.

최근의 연구에서 Ding 외 4인은 클라우드 서비스 영역에서 사용자 유사도에 시간 인지 개념을 적용하여 데이터 희소성 문제 해결과 서비스의 질 향상을 모두 추구한 모델을 제안하였다[23]. Lu 외 3인은 논문 추천을 위하여 시간 인지 신경망의 협력 필터링 시스템을 개발하였는데, 논문의 다차원적인 특성을 고려하여 통합하였다[24]. 또한 추천 시스템의 또 다른 종류인 내용 기반 필터링과 협력 필터링을 결합한 하이브리드 방식에 시간 인지 개념을 접목한 연구 결과도 발표되었다[25]. CF 방식의 두 가지 전통적 분류인 사용자 기반과 항목 기반을 통합하고 이에 시간 인지 모델을 추가하여 추천 알고리즘을 제안한 방법도 개발되었다[26][27].

III. Proposed Methodology

1. Motivation

본 연구에서는 임의의 두 사용자의 공통 평가 항목 수를 평가 시간에 따라 그 중요도를 다르게 부여했을 때, 협력 필터링 시스템에 미치는 성능 변화를 알아본다. 이는 오랜 과거 시간에 공통으로 평가한 항목들의 개수와 최근의 그것들은 두 사용자의 유사도에 미치는 영향이 다를 것이라는 가정을 기반으로 한다. 극단적인 예를 들자면, 사용자 A와 B의 공통 평가 항목 수는 모두 10개이며, 이는 수년 전에 발생하였고, 반면에 사용자 A와 C의 공통 평가 항목 수도 동일하게 모두 10개이며, 최근 한 달 사이에 발생하였다고 하자. 유사도를 공통 평가 항목 수만을 기반으로

산출할 때 A와 B 간의 유사도와 A와 C 간의 유사도 값을 서로 다르게 해야 한다는 것이 본 연구의 기본 아이디어이다. 이 예에서는 A와 C 간의 유사도를 더욱 큰 값으로 결정해야 하며, 이는 A를 위한 추천 리스트를 결정할 때, C의 선호 항목들을 B의 그것들보다 우선순위에 놓아야 한다는 것을 의미한다.

2. Formulation

공통 평가 항목 수에 기반한 유사도를 측정하기 위하여 자카드 계수[9][10]를 도입한다. 두 사용자 u와 v 간의 자카드 계수는 다음과 같이 정의하며 표 1에 사용된 기호를 설명하였다.

$$Jaccard_{u,v} = \frac{|I_u \cap I_v|}{|I_u \cup I_v|}$$

제안 방법에서 시간 인지 협력 필터링을 위한 유사도 산출은 전체 평가 데이터를 평가 시간 구간별로 분할한 후, 각 구간에 해당하는 평가 데이터를 활용하여 유사도를 측정하고, 이와 같이 산출된 각 유사도 값을 통합하는 과정을 거친다. 구체적으로 다음과 같은 절차를 수행한다.

1. 데이터 셋의 전체 평가 시간을 n 개의 구간으로 나눈다. 각 구간을 T_k로 표기하며, T_1은 가장 과거의 구간, T_n은 가장 최근의 구간으로 지정한다.
2. 각 구간 T_k 별로 두 사용자 u와 v 간의 유사도를 산출한다. 유사도는 해당 시간 구간에 이루어진 평가 데이터들만을 활용한다. 구체적으로,

$$sim_{u,v}(T_k) = \frac{|I_u(T_k) \cap I_v(T_k)|}{|I_u(T_k) \cup I_v(T_k)|},$$

$$I_u(T_k) = \{i \in I \mid t(r_{u,i}) \in T_k\}$$

3. 각 구간별로 산출된 유사도를 아래 식을 이용하여 통합함으로써 두 사용자 간의 최종 유사도를 산출한다.

$$sim_{u,v} = \frac{\sum_{k=1}^{n} w(T_k) \cdot sim_{u,v}(T_k)}{\sum_{k=1}^{n} w(T_k)},$$

$$w(T_k) = e^{\lambda k}, \quad \lambda > 0$$

Table 1. Description of symbols

Symbol	Description
I	set of items
$r_{u,i}$	rating value given by user u to item i
$t(r_{u,i})$	rating time of user u for item i
T_k	the kth time interval
I_u	set of items rated by user u
$I_u(T_k)$	I_u during the kth time interval
$w(T_k)$	weight for the kth time interval

3. Description of the Proposed Method

앞 절에서 정의한 두 사용자 간의 최종 유사도 값은 각 시간 구간의 유사도 값의 가중 합으로 산출된다. 기존의 시간 인지 협력 필터링 관련 연구[11][18]에서 과거 평가치에 대해 기하급수적으로 낮은 가중치를 부여한 사실을 토대로, 본 연구에서도 $w(T_k)$ 함수는 최근의 유사도 값 일수록 기하급수적으로 가중하는 지수함수로 정의하였다. 따라서 최근에 두 사용자가 공통으로 평가한 항목수가 많을수록 최종 유사도 값은 급격하게 증가한다.

λ 값은 기하급수적 가중치의 크기를 결정하는 파라미터로서 이 값을 통하여 최근 시간 구간의 유사도가 최종 유사도 값에 미치는 영향을 조정할 수 있다. 지수 함수 값이 k가 커짐에 따라 매우 급하게 상승함을 고려할 때, λ 값의 크기는 상당히 작은 값으로 설정하는 것이 시스템의 성능을 위하여 바람직할 것으로 판단되며, 이는 물론 실험을 통하여 결정하여야 한다.

평가 데이터의 전체 시간을 몇 개의 구간으로 분할하느냐의 문제, 즉, n값의 결정 문제는 시스템 성능을 결정하는 또 다른 중요한 이슈이다. n=1이면 시간 인지가 반영되지 않은 원래의 협력 필터링 시스템이다. n값이 매우 크면 각 시간 구간의 길이는 짧아지므로 각 구간 내의 공통 평가 항목 수는 매우 적어지게 되어, 산출된 유사도의 신뢰가 저하되거나 공통 평가 항목수가 0일 가능성도 있으므로 유사도 산출 자체가 불가하게 될 수 있다. 따라서 최적의 성능을 위한 n값은 평가 데이터의 밀집도, 데이터 셋의 전체 시간 구간 길이 등에 의해 좌우되며 실험에 의해 결정되어야 할 파라미터이다.

IV. Performance Experiments

1. Dataset Description

제안 방법은 각 사용자의 평가치 뿐만 아니라 평가 시간이 필요하므로 개방된 연구용 데이터 셋들 중에서 이러한 정보를 포함한 셋을 활용해야 한다. MovieLens 데이터 셋은 관련 연구에서 널리 유용하게 활용되며 본 연구 방법의 실험 조건에 부합되므로 이를 선정하였다.

MovieLens 제공 사이트에서는 평가 데이터 규모가 100K, 1M 등의 셋을 제공한다. 본 연구에서는 규모가 큰 1M을 선택하였다. 이 데이터 셋에서 각 사용자는 최소 20개 이상의 1~5 사이의 정수 평가치를 부여하였고, 전체 평가 기간은 2000년 이후부터 약 34개월간이다. 각 평가 데이터는 사용자ID, 영화ID, 평가치, 타임스탬프 정보를 갖는다.

공통 평가 항목 수에 의존하는 Jaccard 계수의 특성을 고려하고 제안 방법의 실험을 위하여 다음과 같은 절차로서 실험 데이터를 구축하였다.

1. 각 사용자의 평가개수가 150 이상인 건수들만 추출한다. 이 조건에 부합하는 평가개수는 총 744,072이고, 사용자수는 2,096이다. 결과로서 희소성 수준(sparsity level)은 0.9102로서, 1M의 0.9581 보다 밀집된 결과를 나타냈다.
2. 1의 결과 데이터 셋을 평가 시간별로 그룹화 하였다. 성능 결과를 산출하기 위하여 평가 시간 구간(Time Interval, TI)을 1부터 17 까지 다양화하였다. TI=1은 한 개의 구간이 1 개월분으로 짧으며, 원 데이터의 평가 기간이 34 개월이므로 구간의 총 개수는 34개이다. 반면에 TI=17은 원 데이터의 전체 구간을 반으로 나눈 것으로서, 한 개의 구간은 17 개월이며 구간의 총 개수는 2이다. 이와 같이 각 시간 구간 그룹의 크기 및 개수를 다양화함으로써 성능을 보다 면밀히 검토할 수 있다.

2. Performance Metrics

전체 데이터 중에서 80%를 훈련 데이터로 설정한 후, 이를 이용하여 유사도 측정 방법에 따라 현 사용자의 인접 이웃들을 구한다. 남은 20%의 테스트 데이터를 활용하여 성능 결과를 산출한다. 협력 필터링 연구 분야에서 주로 사용되는 성능 평가 척도들을 도입하였으며 상세 내용은 다음과 같다[3][5].

1. MAE(Mean Absolute Error, 평균절대오차): 미평가 항목에 대하여 시스템에서 예측한 평가치와 실제 평가치와의 평균 차이를 나타낸다. 사용자 u가 항목 x에 대하여 부여한 실제 평가치를 ru,x라고 하고, 이에 대한 시스템의 예측치를 $\hat{r}_{u,x}$라고 할 때 MAE의 정의는 아래와 같다.

$$MAE = \frac{1}{n}\sum_{u}\sum_{x}|r_{u,x} - \hat{r}_{u,x}|$$

2. Precision(정밀도): 시스템이 추천하는 항목들에 대한 사용자의 만족도를 측정한다. 데이터 셋의 1부터 5까지의 평가치 범위를 고려할 때 예측치가 4 이상인 항목들을 추천 리스트에 포함하였으며, 추천 항목들 중에서 실제 평가치가 4 이상인 항목들의 비율을 나타낸다.
3. Recall(재현율): 시스템의 추천의 질을 평가하는 또 다른 척도로서, 실제 평가치가 4 이상인 전체 항목들 중에서 시스템이 예측치를 4 이상으로 산출하여 추천 리스트에 포함시키는 항목들의 비율이다.
4. Coverage(커버리지): 시스템의 항목들 중에서는 현 사용자의 인접이웃들 중 아무도 평가치를 부여하지 않은 항목들이 있을 수 있다. 이러한 개념의 역으로서 커버리지란 전체 항목들 중에서 시스템의 평가 예측치 산출이 가능한 항목들의 비율이며, 예측 정확도와 추천 정확도 외에 많은 추천 시스템에서 사용하는 성능 평가 척도들 중의 하나이다.

3. Results of Experiments

3.1 MAE and Coverage

그림 1은 시간 구간(TI)의 변화에 따른 제안 방법과 Jaccard 방법의 MAE와 coverage 결과이다. 제안 방법은 TAJ(Time-Aware Jaccard)로 표기하였고, λ=0.1을 적용하였다. 전반적으로 시간 구간이 달라짐에 따라 성능 차이가 크게 나타났으므로 제안 방법의 시간 구간 값의 설정이 중요함을 알 수 있다. TI 값이 작으면 하나의 시간 구간에 포함되는 평가 개수도 적은데, 이러한 경우에 각 구간 내에서의 공통 평가 항목 수도 적어지므로, TAJ의 성능이 저하된다. 따라서 그림 1에서 TI=1일 때 성능이 가장 낮고, 그 다음으로 TI=5일 때 낮은 성능을 보였다.

이와는 반대로, TI 값이 큰 경우, 즉, TI=9, 12, 17일 때는 하나의 시간 구간 내 평가 개수가 많아지고 공통 평가 항목 수도 점점 많아지게 된다. 따라서 작은 값의 TI에 대한 성능보다 훨씬 좋은 결과를 나타냄을 그림 1에서 확인할 수 있다. 특히, coverage의 경우엔 다른 방법들에 비해 가장 우수한 성능을 보였다. Jaccard 방법은 큰 값의 TI에 대한 TAJ 실험 성능과 매우 대등한 결과를 보였는데, MAE와 coverage의 두 척도에서 모두 해당됨을 알 수 있다.

TI=7인 경우는 하나의 시간 구간이 7개월의 기간임을 나타내고, 따라서 전체 구간 개수는 34개월/7로서 약 5개이다. 이에 대한 TAJ 성능은 MAE 측면에서 놀랍게도 가장 우수하였는데 Jaccard를 뛰어넘는 결과를 보였다. 이로써, 적절한 범위의 시간 구간을 설정한다면, 시간 인지 방식을 적용하는 제안 방법은 기존의 Jaccard의 예측 정확도를 개선할 수 있음을 입증하였다. 다만 coverage 결과에서는 TI=7의 경우에 가장 우수한 성능의 그룹에 속하지 않는 것으로 나타났다.

3.2 Recommendation Quality

그림 2는 추천의 질을 측정한 결과이다. Precision과 recall의 결과는 서로 비교가 용이하도록 동일한 범위 크기, 즉, 0.25로 설정하였다. 이는 그림 1의 coverage와 동

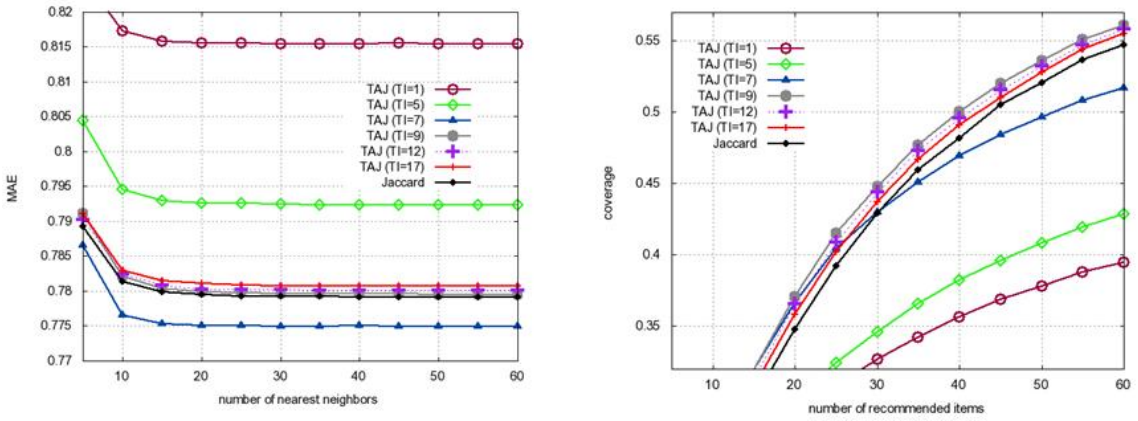

Fig. 1. MAE and coverage performance of the methods

일한 범위 크기이며 MAE의 범위 크기인 0.05 보다는 훨씬 크다.

Precision의 결과를 자세히 살펴보면 실험 방법들 중에서 최고와 최저 간의 성능 차이는 추천 항목 개수가 60일 때 발생하며, 그 차이는 대략 0.067인 반면에 coverage 결과에서는 약 0.166이다. 따라서 실험 방법들 간의 성능 차이가 coverage에서 더욱 크게 발생함을 알 수 있다.

Precision 결과에서 TI=1 보다 TI=5의 경우에 더욱 저조한 성능을 보였는데, 이는 그림 1의 결과와는 반대이다. 추천 성능에서는 정밀한 예측치를 산출하는 데에 초점이 있는 것이 아니라 미평가 항목에 대한 예측치가 정해진 한계선 이상 인지 또는 미만 인지만을 판단하기 때문에, 이와 같이 예상과는 다소 다른 결과가 나타난 것으로 판단된다. Jaccard 방법은 precision 측면에서 예상 밖으로 가장 우수한 그룹들 중 하나에 속하는 것으로 나타났다. 제안방법 실험에서 TI 값이 큰 경우(TI=9, 12, 17)에 작은 경우 보다 우수한 precision 결과를 보임을 알 수 있다. 이러한 발견은 그림 1의 coverage 결과에서도 동일하게 확인된다.

그림 2의 recall에서는 precision과는 다른 양상을 볼 수 있는데, TI=5인 경우에 가장 우수한 성능을 나타내었고, TI=7이 그 다음으로 우수하였다. 나머지 방법들은 거의 유사한 성능 결과를 보임을 알 수 있다. 그러나 방법들 간의 성능 차이가 precision에서보다 작았다. TI 값을 기준으로 볼 때 precision과 recall이 서로 상반된 결과를 보였으므로, 이 둘의 조화평균인 F1 척도를 산출해 보았다. 그 결과, F1 성능은 precision 성능과 동일한 순위로 나타났는데, 즉, TI값이 클 때 가장 우수하고, TI=5 또는 1인 경우에 가장 낮은 성능을 보였다.

실험 결과를 종합해 보면, 시간 인지의 Jaccard 방법은 거의 모든 성능 척도에서 단순 Jaccard 방법을 능가하는 우수한 결과를 보였다. 다만 모든 척도에서 항상 우수한 결과를 가져오는 TI 값은 일정하지 않았다. 즉, 성능 척도의 종류에 따라 가장 우수한 결과의 TI 값은 다르다는 사실을 확인하였다.

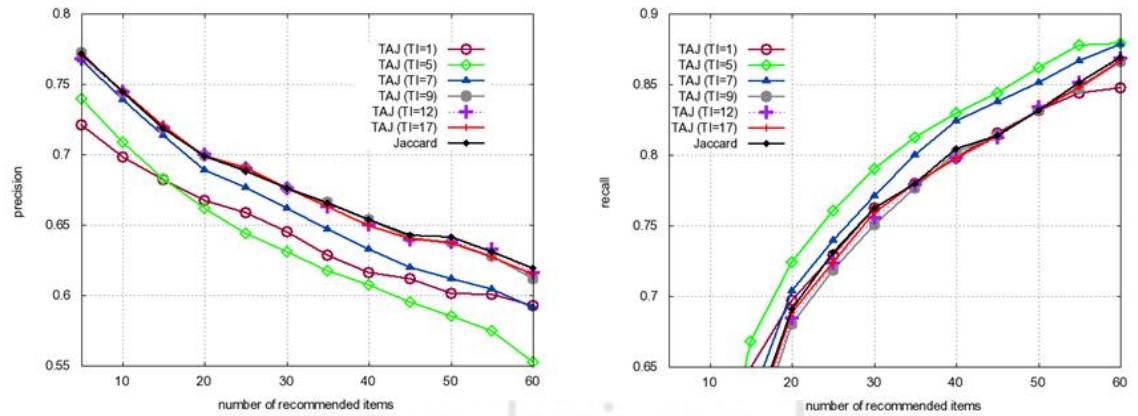

Fig. 2. Recommendation quality of the methods

V. Conclusions

두 사용자 간 공통 평가 항목 집합은 이웃 기반의 협력 필터링 시스템에서 유사도를 산출하기 위한 주요 요소이다. 자카드 계수는 이 집합의 상대적 크기를 측정하며 유사도값에 반영하여 데이터 희소성 문제를 해결하기 위한 목적으로 많은 연구 결과들이 수행되었다.

본 연구에서는 유사도 측정을 위해 두 사용자의 공통 평가 항목수의 비율을 시간에 따라 다른 비중으로 취급하는 것이 시스템의 성능에 어떠한 변화를 가져오는지를 분석하였다. 제안 방법은 시간 인지 개념을 자카드 계수에 적용한 최초의 연구로서, 주기적으로 사용자 간의 자카드 계수를 산출한 후, 이들의 가중 합을 유사도 값으로 정의하여 시스템에 적용한다. 가중치는 시간에 따른 함수값으로 정하고, 시간 주기 변화에 따른 성능 변화를 측정하였다.

다양한 성능 척도를 활용하여 제안 방법의 성능 실험을 수행한 결과, 자카드 계수를 평가 시간에 따라 변환하여 시스템에 도입하는 것이 성능 개선 효과를 가져왔으므로 이와 같은 전략이 반드시 필요하며, 최적의 시간 주기는 성능 척도의 종류에 따라 달라진다는 사실을 확인하였다.

본 연구의 한계점을 열거하면 다음과 같다. 제안 방법은 자카드 계수에 대한 시간 영향을 분석하기 위한 것이므로 실험 비교 대상으로서 자카드 계수만을 포함하였다. 실험을 위한 데이터 셋 구축을 위해 공통 평가 항목수가 매우 적거나 0이 되지 않도록 데이터 희소성 수준을 낮추었는데, 이는 자카드 계수와의 성능 비교가 가능 및 용이하게 하기 위함이었다. 또한 성능 척도의 종류에 따라 제안 방법에 대한 최적의 시간 구간 값은 달라짐을 실험 결과 확인하였는데, 향후 추가 연구 및 실험을 통하여 척도의 종류와 무관한 최적 시간 구간의 제시가 필요할 것이다.

REFERENCES

[1] R. Chen, Q. Hua, Y. -S. Chang, B. Wang, L. Zhang and X. Kong, "A Survey of Collaborative Filtering-based Recommender Systems: From Traditional Methods to Hybrid Methods based on Social Networks," IEEE Access, Vol. 6, pp. 64301-64320, 2018. DOI: 10.1109/ACCESS.2018.2877208

[2] B. Shao, X. Li, and G. Bian, "A Survey of Research Hotspots and Frontier Trends of Recommendation Systems from the Perspective of Knowledge Graph," Expert Systems with Applications, Vol. 165, 2021. DOI:10.1016/j.eswa.2020.113764

[3] M. Jalili, S. Ahmadian, M. Izadi, P. Moradi, and M. Salehi, "Evaluating Collaborative Filtering Recommender Algorithms: A Survey," IEEE Access, Vol. 6, pp. 74003-74024, 2018. DOI: 10.1109/ACCESS.2018.2883742

[4] F. Fkih, "Similarity measures for Collaborative Filtering-based Recommender Systems: Review and Experimental Comparison," Journal of King Saud University - Computer and Information Sciences, Vol. 34, No. 9, pp. 7645-7669, 2022. DOI: 10.1016/j.jksuci.2021.09.014

[5] K. G. Saranya, G. S. Sadasivam, and M. Chandralekha, "Performance Comparison of Different Similarity Measures for Collaborative Filtering Technique," Indian Journal of Science and Technology, Vol. 9, No. 29, 2016. DOI:10.17485/ijst/2016/v9i29/91060

[6] H. Khojamli and J. Razmara, "Survey of Similarity Functions on Neighborhood-based Collaborative Filtering," Expert Systems with Applications, Vol. 185, 2021, Article Number 115482, DOI: 10.1016/j.eswa.2021.115482

[7] H.-F. Sun, et al., "JacUOD: A New Similarity Measurement for Collaborative Filtering," Journal of Computer Science and Technology, Vol. 27, No. 6, pp. 1252-1260, 2012. DOI: 10.1007/s11390-012-1301-5

[8] A. A. Amer, and L. Nguyen, "Combinations of Jaccard with Numerical Measures for Collaborative Filtering Enhancement: Current Work and Future Proposal," ArXiv. /abs/2111.12202, 2021. DOI: 10.48550/arXiv.2111.12202

[9] G. Koutrica, B. Bercovitz, and H. Garcia, "FlexRecs: Expressing and Combining Flexible Recommendations", Proc. the ACM SIGMOD International Conference on Management of Data, pp. 745-758, 2009. DOI: 10.1145/1559845.1559923

[10] S. Lee, "Improving Jaccard Index for Measuring Similarity in Collaborative Filtering," Lecture Notes in Electrical Engineering, Vol 424, 2017. DOI: 10.1007/978-981-10-4154-9_93

[11] P. G. Campos, F. Diez, and I. Cantador, "Time-aware Recommender Systems: A Comprehensive Survey and Analysis of Existing Evaluation Protocols," User Modeling and User-Adapted Interaction, Vol. 24, No. 1, pp. 67-119, 2014. DOI: 10.1007/s11257-012-9136-x

[12] A. Livne, E. S. Tov, A. Solomon, A. Elyasaf, B. Shapira, and L. Rokach, "Evolving Context-aware Recommender Systems with Users in Mind," Expert Systems with Applications, Vol. 189, 2022, Article Number 116042, DOI: 10.1016/j.eswa.2021.116042

[13] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering," ACM SIGIR Forum, Vol. 51, No. 2, pp. 227-234, 2017. DOI: 10.1145/3130348.3130372

[14] H. -J. Kwon, T. -H. Lee, J. -H. Kim, and K. -S. Hong, "Improving Prediction Accuracy using Entropy Weighting in Collaborative Filtering," Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, pp. 40-45, 2009. DOI: 10.1109/UIC-

ATC.2009.50.

[15] H. Liu, Z. Hu, A. Mian, H. Tian, and X. Zhu, "A New User Similarity Model to Improve the Accuracy of Collaborative Filtering", Knowledge-Based Systems, Vol. 56, pp. 156-166, 2014. DOI: 10.1016/j.knosys.2013.11.006

[16] N. Mohammadi and A. Rasoolzadegan, "A Two-stage Location-sensitive and User Preference-aware Recommendation System," Expert Systems with Applications, Vol. 191, 2022, Article Number 116188, DOI: 10.1016/j.eswa.2021.116188

[17] Y. Huai-Zhen and L. Lei, "An Enhanced Collaborative Filtering Algorithm Based on Time Weight," International Symposium on Information Engineering and Electronic Commerce, pp. 262-265, 2009. DOI: 10.1109/IEEC.2009.61

[18] Y. Ding and X. Li, "Time Weight Collaborative Filtering," Fourteenth ACM International Conference on Information and Knowledge Management, pp. 485-492, 2005. DOI: 10.1145/1099554.1099689

[19] L. He and F. Wu, "A Time-Context-Based Collaborative Filtering Algorithm," IEEE International Conference on Granular Computing, pp. 209-213, 2009. DOI: 10.1109/GRC.2009.5255130

[20] N. Zheng and Q. Li, "A Recommender System based on Tag and Time Information for Social Tagging Systems," Expert Systems with Applications, Vol. 38, No. 4, pp. 4575-4587, 2011. DOI: 10.1016/j.eswa.2010.09.131

[21] C. Wangwatcharakul and S. Wongthanavasu, "A Novel Temporal Recommender System based on Multiple Transitions in User Preference Drift and Topic Review Evolution," Expert Systems with Applications, Vol. 185, 2021, Article Number 115626, DOI: 10.1016/j.eswa.2021.115626

[22] G. Xu, Z. Tang, C. Ma, Y. Liu, and M. Daneshmand, "A Collaborative Filtering Recommendation Algorithm Based on User Confidence and Time Context," Journal of Electrical and Computer Engineering, Vol. 2019, Article ID 7070487, DOI: 10.1155/2019/7070487

[23] S. Ding, Y. Li, D. Wu, Y. Zhang, and S. Yang, "Time-aware Cloud Service Recommendation using Similarity-enhanced Collaborative Filtering and ARIMA Model," Decision Support Systems, Vol. 107, pp. 103-115, 2018. DOI: 10.1016/j.dss.2017.12.012.

[24] Y. Lu, Y. He, Y. Cai, and Z. Peng, "Time-aware Neural Collaborative Filtering with Multi-dimensional Features on Academic Paper Recommendation," IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2021. DOI: 10.1109/CSCWD49262.2021.9437673

[25] H. Li and D. Han, "A Time-aware Hybrid Recommendation Scheme Combining Content-based and Collaborative Filtering," Frontiers of Computer Science 15, Vol. 154613, 2021. DOI: 10.1007/s11704-020-0028-7

[26] S. Lee, "Time-aware Collaborative Filtering with User- and Item-based Similarity Integration," Journal of the Korea Society of Computer and Information, Vol.27, No. 9, pp. 149-155, 2022. DOI: 10.9708/jksci.2020.25.12.000

[27] Y. Wan, Y. Chen, and C. Yan, "An Integrated Time-Aware Collaborative Filtering Algorithm," Knowledge Management in Organizations, pp. 369-379, 2021. DOI: 10.1007/978-3-030-81635-3_30

Authors

Soojung Lee received the B.S. degree in Mathematics Education from Ewha Woman's University, Korea in 1985. She received M.S. and Ph.D. degrees in Computer Science from Texas A&M University in 1990 and 1994, respectively. Dr. Lee joined the faculty of the Department of Computer Education at Gyeongin National University of Education, Gyunggi-do, Korea, in 1998, as a professor. She is interested in recommender systems, information filtering, data mining techniques, and computer education.