韓國 컴퓨터情報學會 論文誌
第14卷 第1號, 2009. 1.

2009-14-1-1-1

인터넷 채팅 도메인에서의 감성정보를 이용한 다관점 사용자 선호도 학습 방법

신 욱 현*, 정 윤 재*, 맹 성 현**, 한 경 수***

Multi-perspective User Preference Learning in a Chatting Domain

Wookhyun Shin*, Yoonjae Jeong*, Sung-Hyon Myaeng**, Kyoung-Soo Han***

요 약

개인화 서비스와 같은 지능정보 시스템을 위해서는 사용자 선호도의 학습은 중요한 연구 분야이다. 본 연구에서는 채팅 도메인에서의 사용자 선호도를 학습하는 방법을 제시하며, 기존의 평면적인 사용자 선호도 모델의 문제점을 해결하기 위한 사용자 선호도 모델을 제안한다. 사용자가 선호도 학습의 대상에 대하여 얼마나 관심이 있는가를 나타내는 관심도와 대상에 대한 감성을 나타내는 호감도라는 요소로 모델링 할 수 있다. 자연어 처리를 통해 현재 대화에서의 주제 탐지와 호감도 분석을 하고, 이를 이용하여 사용자의 선호도와 호감도를 학습한다. 시간의 흐름에 따라 변하는 사용자 선호도의 특징을 고려하여, 사용자 선호도를 세션, 단기, 장기 선호도로 나누어 계산한다. 사용자 선호도 학습의 대상이 되는 키워드와 주제에 대하여 시간에 따라 변하는 사용자의 선호도 변화를 고려하여 선호도 결정을 한다. 사용자 선호도 학습 효과의 검증을 위하여 사용자 평가를 하였으며 주제 선호도, 키워드 선호도, 키워드 호감도에 대하여 각각 86.52%, 86.28%, 87.22%의 성능을 보였다.

Abstract

Learning user's preference is a key issue in intelligent system such as personalized service. The study on user preference model has adapted simple user preference model, which determines a set of preferred keywords or topic, and weights to each target. In this paper, we recommend multi-perspective user preference model that factors sentiment information in the model. Based on the topicality and sentimental information processed using natural language processing techniques, it learns a user's preference. To handle time-variant nature of user preference, user preference is calculated by session, short-term and long term. User evaluation is used to validate the effect of user preference learning and it shows 86.52%, 86.28%, 87.22% of accuracy for topic interest, keyword interest, and keyword favorableness.

▸ Keyword : 사용자 선호도(User Preference), 감성 분석(Sentiment Analysis), 사용자 모델(User Model)

• 제1저자 : 신욱현
• 투고일 : 2008. 11. 7, 심사일 : 2008. 12. 5, 게재확정일 : 2008. 12. 28.
* 한국정보통신대학교 공학부 ** 한국정보통신대학교 정교수 *** SK텔레콤
※ 본 연구는 지식경제부 및 정보통신연구진흥원의 IT핵심기술개발사업의 일환으로 수행하였음. [2008-F-047-01, Urban Computing Middleware 기술 개발]

Ⅰ. 서 론

채팅은 인터넷 서비스 중 가장 많이 이용되는 서비스 중의 하나이다. 사용자는 메신저나 채팅 사이트에서의 대화를 통해 자신의 개인적인 관심사와 감정적인 태도를 직접적으로 드러낸다. 따라서, 채팅 상의 대화 내용은 사용자의 선호(preference)를 학습하는데 있어 풍부한 정보를 제공해 준다.

사용자 선호도 학습은 인터넷 정보 서비스의 개인화 관점에서 오랫동안 연구되어 왔다[1-3]. 사용자 선호도 학습 방법으로는 웹 마이닝 (web mining)[4], 협력적 필터링(collaborative filtering)[5], 문서 구조 분석[6] 등의 연구 방향이 존재한다. 최근 정보 검색 분야에서는 사용자의 선호도를 분석하여 검색에 활용하기 위해, 검색 결과에 대한 사용자의 암시적인 반응(implicit feedback)과 클릭 데이터(click-through data)를 분석하는 시도도 이루어지고 있다[5-6].

사용자의 선호도에 감성 정보를 고려한 연구로는 [7]이 있다. [7]은 사용자가 뉴스 포탈 사이트에서 다양한 시각의 기사를 편중됨 없이 읽을 수 있게 하는 것을 목적으로 한다. 예를 들면 이라크 테러라는 주제의 경우, 자살 테러와 같은 슬픈 감성의 기사를 많이 읽는 사용자에게 국제 구호대가 이라크에서 구호 활동을 벌이는 대칭되는 감성 정보를 가지는 기사를 추천해 준다. 이를 위해 특정 주제에 관련된 뉴스 기사를 사용자가 선호하는 감성에 부합하는 주류 기사와 부합하지 않는 비주류 기사로 분류하고, 비주류 기사를 제공하는 기사 제공을 목적으로 한다[7].

기존의 연구들은 사용자의 대화기록, 인터넷 사용 기록 등을 기반으로 주제 혹은 키워드에 대한 사용자의 선호도를 파악한다. 이는 선호도가 뉴스에서 어느 분류의 기사를 많이 읽는가 혹은 어떠한 가수의 노래를 많이 듣는가와 같이 단편적으로 구성됨을 의미한다. 하지만 사용자가 얼마나 관심을 가지는지 여부 만을 고려하는 기존의 사용자 선호도에서는 주제 혹은 키워드에 대하여 어떠한 감성을 갖는지 파악할 수 없다. [7]의 연구 역시 문서의 감성 정보를 활용하지만 어떤 대상에 대한 개인의 태도가 아닌 긍정적인 혹은 부정적 기사에 대한 선호도를 학습할 뿐이다.

본 논문은 인터넷 채팅에서 사용자의 선호도에 대한 새로운 모델과 이를 학습하기 위한 방법론을 제안한다. 제안된 모델은 선호도 대상에 대한 관심도(interest)와 호감도(sentiment)로 구성된다. 관심도란 사용자가 대상에 대해 얼마나 흥미를 가지고 있는가를 말하며, 호감도는 대상에 대한 사용자의 감정적 태도를 의미한다. 관심도를 측정하기 위해 사용자가 언급한 빈도 정보를 사용하며, 호감도를 위해 사용자 발화의 양태(modality)를 실마리(clue)로서 활용한다. 제시된 방법은 또한 시간에 따른 선호도의 변화를 반영하여 현재 사용자의 선호도를 알 수 있도록 고안되었다.

본 논문은 다음과 같이 구성이 된다. 2장에서는 본 연구가 제안하는 사용자 선호도 모델에 대하여 기술하며, 3장에서 사용자 선호도 구축 방법을 설명한다. 4장에서는 사용자 평가를 통하여 사용자 선호도 구축을 분석하고, 5장에서 본 연구의 결론을 맺는다.

Ⅱ. 사용자 선호도 모델

기존의 사용자 선호도 모델은 사용자가 선호하는 대상, 혹은 선호대상에 대한 가중치를 결정 하는 선형적 사용자 선호도 모델을 가지고 있다. 본 논문에서는 사용자 선호도를 관심/호감으로 세분화하며, 시간에 따른 사용자 선호도의 변화와 주제와 키워드에 대해 독립적으로 사용자 선호를 학습한다.

본 논문에서는 사용자의 선호도를 두 개의 관점에서 학습한다. 사용자선호도는 관심도(혹은, 흥미도)와 호감도로 나누어질 수가 있다. 관심도는 사용자가 선호의 대상에 대하여 얼마나 정보를 얻기를 원하는가에 관한 척도이다. 예를 들어, 한반도 대운하와 관련된 이야기를 많이 하는 사용자의 경우 한반도 대운하에 대하여 높은 관심을 보인다고 할 수 있다. 선호의 대상에 대하여 사용자가 좋아하는(호감) 정도에 따른 척도인 호감도는 관심도와 별도로 존재 한다. 예를 들어, 사용자가 한반도 대운하에 대하여 반대(비호감)하는 생각을 가질 수 있다. 관심도와 호감도는 상호 독립적으로 존재한다고 가정하며, 그림1과 같이 도식화 할 수 있다. 선호도 학습의 대상이 되는 키워드나 대화 주제는 선호도 평면상에 위치한다.

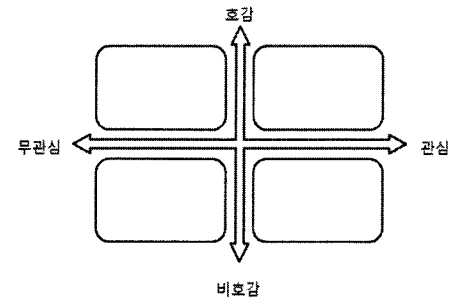

그림 1. 관심도와 호감도로 구성된 사용자 선호도 모델
Fig. 1. Interest-Sentiment User Preference Model

사용자의 관심도는 사용자의 사용 기록에 기반하여 주어진 기간에 출현하는 키워드 및 주제의 빈도를 측정하여 계산한다. 호감도 판단은 고려대학교에서 개발한 양태분석기의 결과에 기반한다. 양태분석기는 입력으로 주어지는 자연어 문장에 대하여 양태를 나타내는 표현이 존재하는지 여부와 구조를 분석하여 문장의 양태를 한 가지로 결정한다.[8]

기존의 선호도 모델은 사용자 관심의 정도만을 제공하기 때문에, 사용자의 감성에 대한 정보는 알 수 없다. 선호도를 이용한 음악 추천이나 상품광고와 같은 어플리케이션의 경우는 사용자의 관심도와 호감도가 모두 높은 대상에 대해 서비스를 제공해야 한다. 한편, 뉴스 추천이나 공익광고와 같은 경우는 사용자와 대칭인 호감도의 서비스를 제공해야 한다. 예를 들어, 미국산 소고기의 안정성을 홍보하는 공익광고는 그에 대해 높은 관심을 가지면서 미국산 소고기에 비호감을 가지는 사용자에게 노출 시키는 것이 더욱 효율적일 것이다. 제안하는 방법의 경우, 사용자가 얼마나 관심을 가지는가에 대한 정보뿐 아니라 사용자가 가지는 감성정보까지 제공하므로, 어플리케이션 수준에서 그 용도에 맞는 사용자의 감성 선호도를 이용 할 수 있다.

사용자 선호도는 상황이나 시간에 따라 변할 수 있다 [9]. 나이가 들어가면서 사용자의 선호도가 취업에서 결혼으로 이동하거나, 올림픽이 시작하면서 스포츠에 대한 선호도가 높아지는 것이 그러한 예시이다. 따라서 시간변화에 따른 사용자의 선호도 변화를 감지하는 것 또한 사용자 선호도 구축에 중요한 역할을 한다.

기존 연구에서는 사용자 선호도 학습의 대상을 주제 수준이나 단어 수준에서 정의한다. [10]은 사용자의 사용기록을 이용하여 사용자가 관심을 가지는 구문을 찾고 그 구문을 계층적으로 군집화하여 UIH (User interest hierarchy)를 구축한다. UIH는 사용자가 관심을 가지는 주제의 계층이며 상위 계층에 존재하는 용어는 일반적인 관심을 나타내는 반면, 세부적인 관심을 나타내는 용어는 하위 계층에 존재한다. UIH구축을 위하여 제안하는 DHC(Divisive graph-based hierarchical clustering method)는 사용자가 관심을 가지는 용어에 대하여, 종료조건을 만족 할 때까지 재귀적으로 자식 클러스터로 나눈다.

기존의 연구에서는 키워드에 대한 선호도와 주제에 대한 선호도가 통합적으로 고려된다. 예를 들면, 기계학습, 자료구조, 알고리즘이라는 키워드에 관심을 보이는 사용자는 컴퓨터 과학이라는 주제에도 관심을 가지는 것으로 판단한다. 하지만, 실제로는 주제 전체가 아닌 특정 키워드에 관심을 가지는 경우도 존재한다. 예를 들어 베이징 올림픽 이후 장미란 선수에 대한 사용자의 관심은 높지만, 역도 전체에 대한 관심은 높지 않을 수 있다. 또한 같은 정치/사회라는 주제에 대하여도 한나라당과 통합민주당에 대하여 서로 다른 감성을 가질 수 있다. 본 선호도 모델에서는 주제와 키워드에 대하여 독립적으로 선호도를 계산하여 이러한 문제를 해결하고자 한다.

III. 사용자 선호도 구축 방법

채팅도메인의 경우는 일반 텍스트에 비해 대화 주제와 흥미의 변화가 뚜렷하기 때문에, 시간에 따른 선호도 변화 고려의 중요성이 더욱 크다. 본 연구에서는 사용자의 선호도를 세션(현재 대화의 시작과 끝), 단기(1개월, 계절단위 등), 장기(전체 기록) 기간으로 나누어 기간별 사용자 선호도를 계산한 다음, 이를 이용하여 전체 사용자 선호도를 계산한다. 세션선호도는 사용자 대화가 이어지는 현재 세션에서 지속되는 사용자의 선호도를 나타내며, 대화주제, 하루 내의 시각 등이 세션 선호도에 영향을 미친다. 단기 선호도는 최근 1개월이나 한 계절과 같이 상대적으로 짧은 시간 동안 지속되는 사용자의 선호도를 가리킨다. 단기 선호도에 영향을 주는 요소로는 계절, 유행, 사회적 사건 등이 있다. 장기 선호도는 사용자 선호도 구축을 시작한 이래 모든 사용 기록을 기반으로 구축한 사용자 선호도를 의미한다. 전체 사용자 선호도는 기간별 사용자 선호도의 가중치 합으로 결정한다.

그림2는 시간에 따른 사용자 선호 모델을 표현한다. 시간축에서의 한 평면은 사용자 선호도를 나타내며, 사용자 선호도는 관심도와 호감도 축으로 표현이 가능하다. 선호도 학습의 대상이 되는 키워드나 대화 주제는 선호도 평면 상에 위치한다.

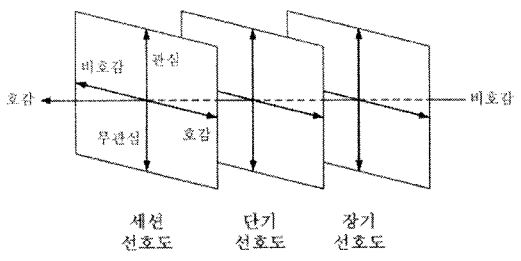

그림 2. 시간에 따른 사용자 선호도 모델
Fig. 2. Timely Variant User Preference Model

선호도 학습의 대상은 키워드와 대화의 주제로 하며, 대화의 주제 탐지를 위해서는 키워드 spotting기법을 이용한다. 주제분류는 생활/건강, 엔터테인먼트, 로맨스/성, 모바일 서비스, 음악을 비롯하여 총 13가지를 사용한다.

세션 선호도는 현재 대화가 지속되고 있는 세션 내에서의 사용자의 선호도이다. 세션 선호도는 매 대화마다 서로 다르게 생성되고 소멸되며, 대화 주제와 밀접하게 연관이 있다. 예를 들어, 사용자가 현재 영화에 대하여 이야기 하고 있다면 사용자의 선호도는 영화에 있다고 생각 할 수 있다. 대화가 지나치게 길어지거나 주제가 산만한 경우는 세션 선호도가 제대로 구축되지 않거나 흐려질 수 있다.

단기 선호도는 1개월에서 몇 달간의 사용자 기록에 기반한 선호도이다. 세션 선호도가 현재 대화에서 사용자의 선호에 대한 정보인 반면, 단기 선호도는 사용자의 최근 관심사에 대한 정보를 제공하며 외부의 변수에 따라 변할 수 있다. 예를 들어 사용자가 최근 자전거 타기라는 취미를 새로 가지게 되어 최근 몇 개월간 기록에 자전거에 대한 이야기가 많다면, 사용자의 단기 선호도는 자전거에 있다고 할 수 있다. 단기 선호도에 영향을 미치는 외부 변수로는 계절, 유행, 사회적 사건, 사용자의 사회적 위치 등이 있다. 계절이 단기 선호도에 영향을 주는 예시로는 여름철에 휴가, 피서에 높은 선호도를 가지는 경우나 겨울에 크리스마스, 해돋이 등에 사용자 선호가 있는 현상 등이 있다. 최근 유행하는 댄스그룹에 선호를 가진다거나, 사회적으로 이슈가 되고 있는 경제 문제 등에 대한 사용자 선호가 있는 것은 유행이나 사회적 사건이 단기 선호도에 영향을 주는 예시이다.

단기 선호도는 외부 변수에 따라 시계열적으로 변할 수 있는 반면, 사용자의 전체 사용 기록으로 계산하는 장기 선호도는 사용자 고유의 가치관을 반영하는 선호도로 정의 한다. 장기 선호도의 예시로는 연예 주제에 대한 비선호, 특정 스포츠 종목에 대한 선호 등이 있다. 세션, 단기 선호도와는 다르게 장기 선호도는 한 번 구축되면 쉽게 변하지 않는 반면 구축하는데 상대적으로 오랜 기간이 소요된다.

3.1 세션 선호도

세션 관심도는 대화 세션 동안 사용자가 키워드에 대해 가지는 태도를 학습하며, 여기서 학습된 결과는 단기 관심도(short-term sentiment)를 계산하기 위한 기반 데이터로 활용된다. 키워드에 대한 호감/비호감은 고려대 양태분석기의 결과를 사용한다. 호감 (LIK)이라고 판명된 발화 내에서 발생한 키워드일 경우, 키워드에 대해 호감이 있는 것으로 간주하며, 비호감 (HAT) 이라고 판명된 발화 내에서 발생한 키워드의 경우, 키워드에 대해 비호감인 것으로 판단한다. 호감도의 값은 0.5를 기준으로 그 이상일 경우 호감을 가진 것으로, 미만일 경우 비호감을 가진 것으로 계산된다.

다음 수식 (3.1)은 키워드에 대한 세션 호감도를 학습하는 방법을 표현한다. $sentiment_{session}(k)$는 키워드 k에 대한 세션 호감도를 나타내며, $freq_{lik}(k)$는 양태분석기에서 호감 (LIK)이라고 판단된 발화 내에서 키워드 k가 발생한 빈도 수, $freq_{hat}(k)$는 비호감 (HAT) 이라고 판단된 발화 내에서 키워드 k가 발생한 빈도수를 나타낸다.

$$sentiment_{session}(k)=\begin{cases}0.5+\frac{0.5}{1+e^{-freq_{lik}(k)}}-\frac{0.5}{1+e^{-freq_{hat}(k)}}, & freq(k)>0\\ 0.5, & freq(k)=0\end{cases} \quad \cdots\cdots (3.1)$$

3.2 단기 선호도

단기 관심도는 1달 동안 사용자가 관심을 표명한 키워드, 대화주제의 관심도를 의미하며, 1달 동안 학습된 세션 관심도의 평균으로 계산된다. 다음 수식 (3.2)에서 $interest_{short-term}(k)$는 키워드 k의 단기 관심도를 나타내며, $interest_{short-term}(t)$는 대화주제 t의 단기 관심도를 의미한다. $N_{short-term}$은 1달 동안 발생한 대화 세션의 수를 가르킨다. 또한, $interest_{i,session}(k)$는 1달 동안의 i번째 세션의 키워드 k에 대한 세션 관심도를 $interest_{i,sessoin}(t)$는 대화주제 t의 i번째 세션의 세션 관심도를 나타낸다.

$$interest_{short-term}(k)=\frac{1}{N_{short-term}}\sum_{i=1}^{N_{short-term}} interest_{i,session}(k)$$
$$interest_{short-term}(t)=\frac{1}{N_{short-term}}\sum_{i=1}^{N_{short-term}} interest_{i,session}(t) \quad \cdots\cdots\cdots\cdots (3.2)$$

단기 호감도는 1달 동안 사용자가 호감/비호감을 표명한 키워드의 호감도를 의미하며, 1달 동안 학습된 세션 호감도의 평균으로 계산된다. 수식 (3.3)에서 $sentiment_{short-term}(k)$는 키워드 k의 단기 호감도를 나타내며, $N_{short-term}$은 1달 동안 발생한 대화 세션의 수를 가르킨다. $sentiment_{i,session}(k)$는 1달 동안의 i번째 세션의 키워드 k에 대한 세션 호감도를 의미한다.

$$sentiment_{short-term}(k)=\frac{1}{N_{short-term}}\sum_{i=1}^{N_{short-term}} sentiment_{i,session}(k) \quad \cdots\cdots\cdots (3.3)$$

3.3 장기 선호도

장기 관심도는 대화 서비스를 사용한 이래로 사용자가 관심을 표명한 키워드, 대화주제의 관심도를 의미하며, 현재까지의 단기 관심도의 평균으로 계산된다. 다음 수식 (3.4)에서 $interest_{long-term}(k)$는 키워드 k의 장기 관심도를 나타내며, $interest_{long-term}(t)$는 대화주제 t의 장기 관

심도를 의미한다. $N_{long-term}$은 현재까지의 단기 기간 (월)의 수를 의미한다. 또한, $interest_{i,short-term}(k)$는 i번째 달의 키워드 k에 대한 단기 관심도를 $interest_{i,short-term}(t)$는 대화주제 t_k의 i번째 달의 세션 관심도를 나타낸다.

$$interest_{long-term}(k)=\frac{1}{N_{long-term}}\sum_{i=1}^{N_{long-term}} interest_{i,short-term}(k)$$
$$interest_{long-term}(t)=\frac{1}{N_{long-term}}\sum_{i=1}^{N_{long-term}} interest_{i,short-term}(t) \quad \cdots\cdots (3.4)$$

장기 호감도는 대화 서비스를 사용한 이래로 사용자가 호감/비호감을 표명한 키워드의 호감도를 의미하며, 현재까지의 단기 호감도의 평균으로 계산된다. 다음 수식 (3.5)에서 $sentiment_{long-term}(k)$는 키워드 k의 장기 호감도를 나타내며, $N_{long-term}$은 현재까지의 단기 기간 (월)의 수를 가르킨다. $sentiment_{i,short-term}(k)$는 1달 동안의 i번째 달의 키워드 k에 대한 단기 호감도를 의미한다.

$$sentiment_{long-term}(k)=\frac{1}{N_{long-term}}\sum_{i=1}^{N_{long-term}} sentiment_{i,short-term}(k) \quad \cdots\cdots (3.5)$$

3.4 최종 선호도 계산

최종 선호도는 사용자의 장기 선호도 (관심도, 호감도)를 기반으로 한다. 그러나, 사용자는 자신이 최근의 관심사에 보다 더 많은 관심을 기울이는 경향이 있다. 따라서, 사용자의 최종 선호도를 계산하기 위해서는 장기 선호도를 기반으로 하되, 가장 최근의 단기 선호도의 값을 가중 시켜주어야 한다.

$$interest(k)=\alpha\times interest_{short-term}(k)+(1-\alpha)\times interest_{long-term}(k)$$
$$sentiment(k)=\alpha\times sentiment_{short-term}(k)+(1-\alpha)\times sentiment_{long-term}(k) \quad \cdots (3.6)$$
$$interest(t)=\alpha\times interest_{short-term}(t)+(1-\alpha)\times interest_{long-term}(t)$$

위 수식 (3.6)은 키워드 k와 대화주제 t에 대한 사용자의 최종 선호도를 의미한다. 단기 선호도에 대한 가중치로 상수 α를 부여하였으며 최종 키워드에 대한 관심도 interest(k)와 호감도 sentiment(k), 그리고 대화주제에 대한 관심도 interest(t)는 이와 같이 계산된다.

IV. 실험 및 결과 분석

제안된 사용자 선호도 학습 기술의 유용성을 확인하기 위해 다음과 같은 실험을 수행하였다. 실험은 20대 성인 7명 (남성 5인, 여성 2인)을 대상으로 하였으며, 2007년 12월 2일부터 8일 까지 1주일간의 사용자 단기 선호도를 실험하였다.

실험은 각 사용자가 1일 1개의 대화 세션을 입력하게 하였으며, 한 세션 당 7~9개 사이의 발화를 입력하도록 하였다. 이를 기준으로 각 사용자의 단기 선호도를 계산하였으며, 계산 결과 나타난 대화주제, 키워드에 대한 단기 선호도 (관심도, 호감도)를 실제 사용자가 동의 하는지를 확인하는 방법으로 수행되었다. 다음 표1은 단기 선호도 계산 결과와 실제 사용자가 이 결과를 동의하는지에 대한 예이다. 예를 들어 첫 번째 행의 의미는 시스템이 생활/건강 주제에 대한 관심도를 0.28로 계산하였으며, 이를 기준으로 선호라고 판단을 하였으나, 이에 대하여 사용자에게 생활/건강 주제에 실제로 관심이 있는지를 물었을 때 사용자는 동의하지 않은 사실을 나타낸다.

표 1. 사용자 선호도 계산 결과와 사용자 동의 여부
Table 1. The Result of Sentiment Preference Learning and User Agreement

대화주제	선호도 계산 결과	선호 여부	사용자 동의 여부
생활/건강	0.28	선호	X
엔터테인먼트	0.32	선호	O
성/로맨스	0.08	비선호	O

4.1 대화 주제 관심도

실험 결과 대화주제에 대한 사용자의 관심도가 0.25 이상일 경우, 사용자가 해당 대화주제에 관심을 가지고 있을 것으로 하였다. 0.25 값은 학습 결과를 보고 실험 수행자가 판단하였다. 실험 결과, 평균 86.52%의 정확도를 보이고 있었다. 그러나, 사용자가 다양한 키워드를 구사하는 경우 (사용자 7), 69.23%로 다른 사용자에 비해 낮은 정확도를 보이고 있었다. 사용자 7의 경우 다른 사용자에 비해 한 문장에 많은 수의 가지고 있다. 현재 사용중인 대화 주제 탐지는 문장이 길어지고 문장 내에 다양한 키워드가 출현하는 경우 낮은 성능을 보인다. 잘못된 주제 탐지로 인하여, 선호도 학습이 이루어 지는 것을 관찰 할 수 있었다. 이 것은 다양한 키워드를 구사하는 경우, 대화주제 탐지의 정확도가 떨어지는 것이 원인으로 판단된다.

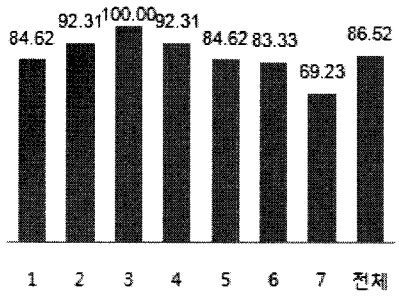

그림 3. 대화주제 관심도 평가 결과
Fig. 3. The Result of Topical Interest Learning

4.2 키워드 관심도

실험 결과 키워드에 대한 사용자의 관심도가 0.42 이상일 경우, 사용자가 해당 키워드에 관심을 가지고 있을 것으로 하였다. 0.42 값은 학습 결과를 보고 실험 수행자가 판단하였다. 실험 결과, 평균 86.28%의 정확도를 보이고 있었다. 그러나, 유사한 개념에 대하여 다양한 표현을 구사하는 경우(사용자 1), 70.00%로 다른 사용자에 비해 낮은 정확도를 보이고 있었다. 사용자1은 "이력서", "졸업", "취직"이라는 비슷한 개념에 대하여 다양한 방법으로 표현하여 개별 키워드의 출현 빈도가 낮기 때문에 키워드 관심도 학습에 실패 하였다.

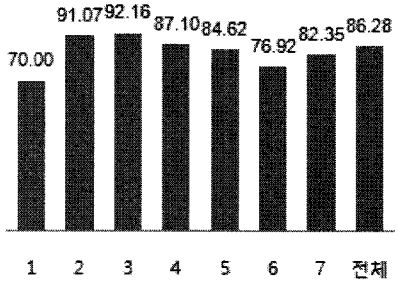

그림 4. 키워드 관심도 평가 결과
Fig. 4. The Result of Keyword Interest Learning

4.3 키워드 호감도

실험 결과 키워드에 대한 사용자의 호감도가 0.6 이상일 경우 호감있는 항목으로, 0.3이하면 비호감인 항목으로 간주하였다. 키워드에 대한 호감도가 0.3초과 0.6미만인 경우는 사용자의 호감/비호감이 뚜렷하지 않다고 생각하여, 이에 대해서는 중립으로 판단하였다. 0.6과 0.3 값은 학습 결과를 보고 실험 수행자가 판단하였다. 실험 결과, 평균 87.22%의 정확도를 보이고 있었다. 대체적으로 높은 정확도를 보이고 있으나, 일부 관용적인 표현에 있어서 오류가 발생하였다 (사용자 2), 이러한 표현으로는 "머리 좋아", "기분 좋아" 등이 있으며, 사용한 양태분석 라이브러리의 오류에 기인한다. 양태분석 라이브러리가 관용적 표현을 포함하는 문장의 양태 판단에 실패하면, 머리, 기분과 같은 키워드에 대하여 잘못된 호감 판단을 하게 된다. 비슷한 예시로는 머리, 기분, 분위기, 기억력, 효과 등이 존재한다.

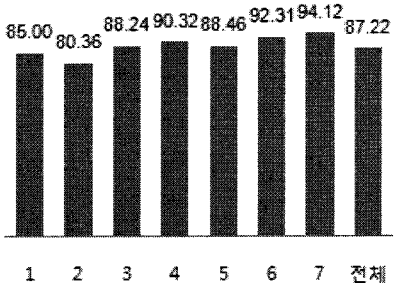

그림 5. 키워드 호감도 평가 결과
Fig. 5. The Result of Keyword Sentiment Learning

본 연구는 시간에 따른 선호도의 변화를 고려한 모델을 제안하고 그에 따른 선호도 계산을 수행한다. 따라서 기간별 선호도를 독립적으로 사용했을 때와의 직접 비교는 수행하지 않는다. 필요한 경우는 사용자의 대화 기록을 제한하고 최종 선호도 계산에서 가중치를 변경하면 세션, 단기, 장기 선호도에 대한 독립적인 평가도 가능하다.

V. 결론

대화도메인을 기반으로 개발된 사용자 선호도 학습 기술은 85% 이상의 높은 정확도를 나타내고 있었다. 하지만 대화주제 관심도의 경우, 다양한 키워드를 사용자가 발화에서 구사할 경우, 상대적으로 낮은 정확도를 보이고 있었으며, 키워드의 관심도 역시 유사한 의미의 다양한 어휘를 구사할 경우 상대적으로 낮은 정확도를 보이고 있었다. 또한 일부 관용적인 표현에 있어서 키워드 호감도를 계산하는데 문제점을 보이고 있었다.

대화주제 선호도의 오류와 키워드 호감도의 오류는 각각 대화주제 탐지 모듈과 양태분석 라이브러리의 성능이 향상될 경우, 같이 성능의 높아질 것으로 기대되지만, 키워드의 관심

도에서 유사한 의미의 다양한 어휘를 고려할 경우는 추가적인 개선 방안이 요구된다. 유사한 개념의 단어에 대해서는 서로와의 연관도를 설정하고, 한 단어가 발생할 경우 유사한 다른 단어의 관심도에도 추가적인 가중치를 주는 방안이 요구된다.

실제 사용자는 시스템을 사용하면서 명시적인 반응을 보이지 않는다. 따라서 본 논문의 실험방법은 대중을 상대로 수행하는 데는 어려움이 있다. 이 문제를 해결하기 위해서는 사용자의 암시적 반응을 통한 평가나 광고, 음악추천과 같은 어플리케이션 이용할 수 있다. 특별히, 장기 선호도의 경우는 시간에 따라 변하지 않는 정도(invulnerable)를 평가 요소로 적용 할 수 있을 것으로 생각된다.

호감도 분석은 양태 분석기의 결과에 기반하는데, 한국어가 아닌 다른 언어에 대해서는 현재 구축되어 있는 양태 분석기를 적용할 수 없다. 따라서 본 방법은 한국어 채팅 도메인에만 국한되는 한계가 있다.

참고문헌

[1] A. Kobsa, "User Modeling: Recent Work, Prospects and Hazards," Human Factors In Information Technology, vol. 10, pp. 111-111, 1993.

[2] 김성희, 김수형, "수준별 동적 교수.학습 시스템 개발을 위한 학습자 모델링 기법," 한국컴퓨터정보학회논문지, 제 7권, 제 2호, 59-67쪽, 2002년 1월.

[3] 한승현, 임영환, "키워드 분석을 이용한 개인화 모바일 웹 뉴스 컨텐츠 생성에 관한 연구," 한국컴퓨터정보학회 논문지, 제 12권, 제 3호, 277-285쪽, 2007년 7월.

[4] S. Holland, M. Ester, and W. KieBling, "Preference Mining: A Novel Approach on Mining User Preferences for Personalized Applications," Lecture Notes In Computer Science, pp. 204-216, 2003.

[5] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative Filtering Recommender Systems," Lecture Notes In Computer Science, vol. 4321, p. 291, 2007.

[6] Q. Feng and C. Junghoo, "Automatic identification of user interest for personalized search," in Proceedings of the 15th international conference on World Wide Web Edinburgh, Scotland: ACM, 2006.

[7] Y. Kawai, T. Kumamoto, and K. Tanaka, "Fair News Reader: Recommending News Articles with Different Sentiments Based on User Preference," Lecture Notes In Computer Science, vol. 4692, pp. 612-622, 2007.

[8] Min Jeong Kim, Sang-Bum Kim, Kyoung-Soo Han and Hae-Chang Rim, "Modality Analysis for Spoken Language Processing," The First Europe-Korean Workshop on Spoken Dialog System Technology , 2008

[9] C. Elisabeth and V. Manuela, "Learning Dynamic Preferences in Multi-Agent Meeting Scheduling," in Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology: IEEE Computer Society, pp. 487-490, 2005.

[10] H. R. Kim and P. K. Chan, "Learning Implicit User Interest Hierarchy for Context in Personalization," Applied Intelligence, vol. 28, pp. 153-166, 2008.

저 자 소 개

신 욱 현

2007년: 한국정보통신대학교 전산학 학사.
현재: 한국정보통신대학교 전산학과 석사과정.
관심분야: Blog Search, Trend Analysis, Social Intelligence, Ads Placement.

정 윤 재

1998년: 포항공과대학교 전산학 학사.
2007년: 한국정보통신대학교 전산학 석사.
현재: 한국정보통신대학교 전산학과 박사과정.
관심분야: Knowledge Discovery, Social Intelligence, Complex System.

맹 성 현

1983년: California State University, Hayward, 전산학 학사.
1985년: Southern Methodist University, Dallas, Texas, 전산학 석사.
1987년: Southern Methodist University, Dallas, Texas, 전산학 박사.
현재: 한국정보통신대학교 공학부 정교수.
관심분야: Information Retrieval, Text Mining, Natural Language Processing.

한 경 수

1998년: 고려대학교 컴퓨터학과 학사.
2000년: 고려대학교 컴퓨터학과 석사.
2006년: 고려대학교 컴퓨터학과 박사.
2006년: 고려대학교 컴퓨터정보통신연구소 연구조교수.
현재: SK 텔레콤 재직.
관심분야: Information Retrieval, Text Mining, Natural Language Processing.