한국지능정보시스템학회논문지 제10권 제3호 2004년 12월(pp.133~146)

사용자 지식을 반영한 메일 폴더 추천 방법론

류미
경희대학교 박사과정, 한국NCR
(ubeauty@hanmail.net)

박주석
경희대학교 교수
(jspark@khu.ac.kr)

김재경
경희대학교 교수
(jaek@khu.ac.kr)

네트워크 기술의 발달로 인하여 사용자가 접하게 되는 정보의 종류와 양이 급속하게 증가되고 있으며, 이로 인해 사용자는 자신이 필요로 하는 정보를 찾아내어 관리하는데 많은 시간과 노력을 소비하고 있다. 이에 본 연구에서는 대표적인 추천기법 중에 하나인 내용기반 추천(Content-based Recommendation)과 사용자 지식에 의해 정의된 키워드 유사성(Keyword Affinity)을 이용하여 사용자가 보다 적은 비용으로 자신의 정보를 효율적으로 관리할 수 있도록 지원하는 방법론을 제시한다. 즉, 사용자의 선호도가 자주 변하거나 새로운 내용이 지속적으로 생성되는 환경에서는 추천의 성능이 떨어지고, 사용자의 선호도가 충분히 축적되기까지 정확한 추천이 어려운 내용기반 추천의 한계점을 사용자 지식에 의해 정의된 키워드 유사성을 응용하여 해결한다.

본 연구는 수시로 새로운 정보가 생성되고 삭제되는 개인 이메일 환경을 그 대상으로 하며, 사용자의 효율적인 이메일 관리를 위한 폴더 추천을 지원한다. 또한 실험을 통해 기존에 연구되었던 폴더 추천 방법론과 성능을 비교함으로써 본 연구에서 제시하는 방법론을 검증하였다.

Key words : 추천시스템, TF-IDF, 사용자 지식, 키워드 유사성(keyword affinity)

논문접수일 : 2004년 11월　　게재확정일 : 2004년 12월　　교신저자 : 김재경

1. 서론

네트워크 기술의 발달로 인하여 사용자들은 다양하고 많은 양의 정보를 접하게 되었다. 이로 인해 사용자들은 수많은 정보 중에서 자신이 필요로 하는 정보를 찾아내기까지 많은 시간과 노력이 필요하게 되었으며, 투자되는 시간과 노력을 줄이기 위한 다양한 연구들이 진행되어 왔다(Belkin and Croft, 1992; Basu et al. 2001; Cohen, 1996; Foltz and Dumaiz 1992). 이러한 연구들 중 내용기반 추천 (Content-based Recommendation)은 텍스트 기반의 아이템부터 영화나 음악에 대한 추천까지 다양한 분야에서 활용되고 있다(Balabanovic and Shoham, 1997; Schafer et al. 2001).

정보 검색(Information Retrieval)과 정보 필터링(Information Filtering)으로부터 발전된 내용기반 추천은 아이템의 속성과 사용자의 선호도간의 유사도를 측정하여 유사도가 높은 순으로 추천 아이템을 결정하는 방법이다. 하지만 이러한 추천 방식은 사용자의 과거 선호도를 학습하여 추천 아이템을 결정하기 때문에, 선호도가 자주 변하거나 새로운 내용이 지속적으로 생성되는 환경에서는 추천의 성능이 떨어지며 다른 영역에 대한 추천이 어렵다(Balabanovic and Shoham, 1997). 또

한 안정적인 추천을 위해서는 충분한 양의 선호도 정보를 학습할 필요가 있다. 따라서 본 연구에서는 내용기반 추천의 한계점을 해결하기 위하여 사용자 지식에 의해 정의된 키워드 유사성(Keyword Affinity)을 응용한다. 키워드 유사성은 하나의 아이디어를 보다 정확하게 표현할 수 있는 방법으로 본 연구에서는 사용자의 선호도를 암시적으로 학습하는데 사용된다.

또한 본 연구에서는 인터넷 사용이 일반화됨에 따라 가장 중요한 커뮤니케이션 방법으로 자리잡은 이메일 환경을 연구대상으로 한다. 많은 양의 신규 이메일이 도착했을 때, 이메일이 포함하고 있는 내용과 관련된 폴더만을 추천해주는 KARS(Keyword Affinity based Recommender System)을 개발하였다. KARS는 사용자의 효율적인 이메일 관리를 지원한다. 본 연구에서는 TF-IDF 가중치를 기반으로 폴더를 추천해주는 SwiftFile 방법론(Segal and Kephart, 2000) 과 사용자 지식과 TF-IDF 가중치를 결합한 KARS 방법론을 비교함으로써 본 연구에서 제안하는 방법론의 우수성을 검증한다.

본 연구의 구성을 보면 다음과 같다. 먼저 2장에서 관련연구와 이론들에 대해 간단히 살펴보고자 한다. 다음으로, 3장에서는 세부적인 알고리즘에 대해 살펴볼 것이며, 4장에서는 실제 이메일 데이터를 이용하여 실험하고자 한다. 마지막으로 5장에서는 본 연구의 결론과 향후 연구 방향에 대하여 논하고자 한다.

2. 문헌연구

2.1 추천시스템과 내용기반 추천

추천시스템은 통계적 기법과 지식탐사 기술(Knowledge Discovery Technology)을 이용하여 고객의 요구에 적합한 아이템을 추천해주는 시스템으로, 고객의 편의를 도모하고 교차판매(cross selling) 및 매출 증대에 초점을 맞춘 시스템이다(Sarwar, et al., 2000). 이러한 시스템은 개인이 처리하기에 너무 많은 양의 정보가 산재해 있는 온라인 환경에서 유용하며, 그 접근 방식은 크게 내용기반 추천(Contents-based Recommendation)과 협업필터링(Collaborative Filtering)으로 나누어 볼 수 있다(Billsus and Pazzani, 1998; Balabanovic and Shoham, 1997). 내용기반 추천은 아이템의 속성에 대한 해당고객의 선호도를 이용하여 추천 아이템을 결정하며, 협업필터링은 고객들의 아이템에 대한 평가를 이용하여 해당고객과 취향이 유사한 고객들이 선호하는 상품을 추천한다(Balabanovic and Shoham, 1997).

정보 검색(Information Retrieval)과 정보 필터링(Information Filtering)으로부터 발전된 내용기반 추천은 주로 텍스트 기반의 아이템에 많이 적용되었으며 그 예로는 InfoFinder(Krulwich, 1996), NewsWeeder(Lang, 1995), WebHount (Chesnais, 1995)가 있다.

본 연구에서는 신규 이메일에 가장 적합한 폴더를 추천하기 위하여 폴더 내에 포함된 이메일의 속성을 분석하는 내용기반 추천 방식을 응용한다.

내용기반 추천을 설계하는데 있어서 아이템에 대한 속성 표현(Feature Representation)이 우선 고려되어야 한다. 내용기반 추천 방식에서 추천 아이템에 대한 속성은 중요한 단어로 표현된다(Pazzani, 1999). 예를 들어 Fab에서는 TF-IDF가중치의 값이 가장 높은 100개의 단어로 문서를 표현했으며, Syskill와 Webert(Pazzani and Billsus,

1997)는 관련 문서를 가장 잘 설명하는 128개의 단어로 문서를 표현했다.

본 연구에서는 각 폴더의 속성을 표현하는데, 일반적으로 잘 알려진 TF-IDF 가중치를 사용하였다. TF-IDF는 해당 단어의 상대적 중요도인 TF(Term Frequency) 와 전체 문서 내에서 해당 단어의 중요도를 나타내는 IDF(Inverse Document Frequency)의 곱으로 다음과 식(1)같이 표현된다 (Salton and McGill, 1983).

$$w_{ij} = f_{ij} \times \log \frac{N}{n_i} \qquad (1)$$

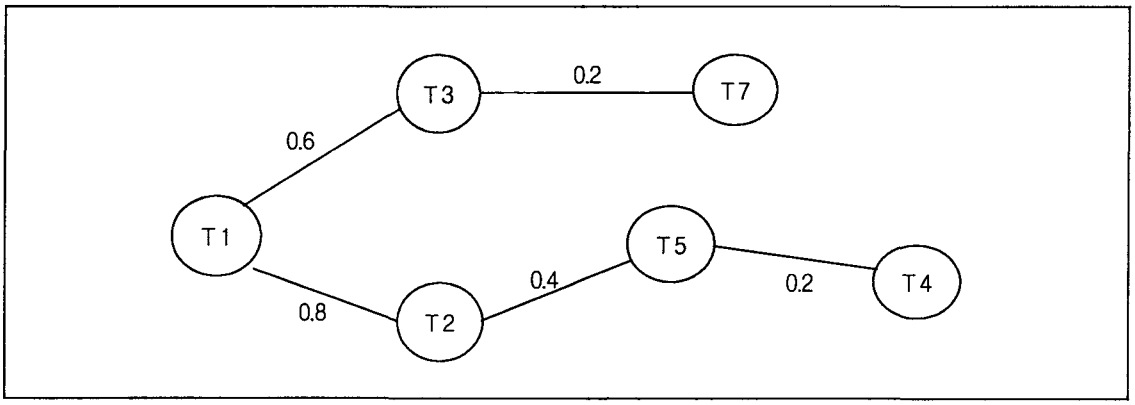
w_{ij} 는 j번째 문서에 나타나는 단어 i에 대한 TF-IDF 가중치이며, 여기서 N은 전체 문서의 수이며, n_i 는 단어 i가 출현한 문서이다. TF-IDF는 특정 단어의 빈도수와 특정 단어가 나타나는 문서의 빈도수를 이용하여 각 단어에 대한 가중치를 구함으로써, 그 단어가 해당 도메인의 특징이나 성격을 얼마나 잘 반영하였는가를 나타내는 방법이다.

아이템에 대한 속성표현이 완성되면 사용자는 추천된 문서에 대한 관심의 정도를 암묵적으로나 명시적으로 표현하게 되며, 시스템은 사용자의 선호도를 계속 누적하여 갱신하는데 이를 사용자 프로파일이라 부른다. (Pazzani, 2000; Kuflik and Shoval, 2000).

본 연구의 폴더 프로파일은 앞서 설명한 사용자 프로파일과 유사한 개념으로, 해당 폴더에 이메일이 포함되거나 삭제될 때마다 관련된 단어의 TF-IDF가중치를 갱신함으로써 폴더의 속성을 항상 최신으로 유지한다.

2.2 키워드 유사성

키워드 유사성(Keyword Affinity)은 유의한 키워드들의 상관관계를 나타내는 것으로 그룹의사결정 과정에서 공통의 의견을 도출하는데 사용되었으며, 고객의 인바운드 이메일에 대한 적합한 응답메일을 찾기 위해 응용되었다(Lee et al. 2002, Yoon et al. 2003). 선행된 연구에 따르면, 지식베이스 안에는 키워드간의 유사성은 다음과 같이 정의되어 있다. 즉 키워드 T1이 있을 때, T1



[그림 1] 키워드간의 유사성

과 T2 그리고 T3 사이에는 각각 0.8과 0.6의 관계가 존재하는 것으로 [그림 1] 과 같이 나타난다.

아이디어의 초기 프로파일이 (T1, T2, T3, T4, T5…Tn) = (1, 0, 1, 0, 1,…1) 일 때, 지식베이스에 있는 키워드간의 유사성을 반영하면 아이디어의 초기 파일은 (T1, T2, T3, T4, T5…Tn) = (1, 0.8, 1, 0, 1,…1)으로 변경된다. 이러한 키워드 유사성은 하나의 아이디어를 보다 정확하게 표현할 수 있는 방법이다(Lee et al. 2002).

본 연구에서는 내용기반 추천의 한계점을 보완하기 위하여 키워드 유사성을 응용하였다. 즉, 내용기반 추천은 이미 축적된 선호도를 학습하여 추천 아이템을 결정하기 때문에 새로운 정보가 수시로 생성되고 삭제되는 이메일 환경에 바로 적용할 수 없으며, 다른 주제의 폴더를 추천하기 어렵다. 또한 선호도 정보가 충분히 축적되지 않은 초기에는 추천의 질이 매우 낮다. 따라서 키워드 유사성을 이용하여 신규 이메일에 대한 속성과 폴더 프로파일의 특징을 보다 정확하게 표현하면 초기부터 보다 정확한 추천을 안정적으로 할 수 있다.

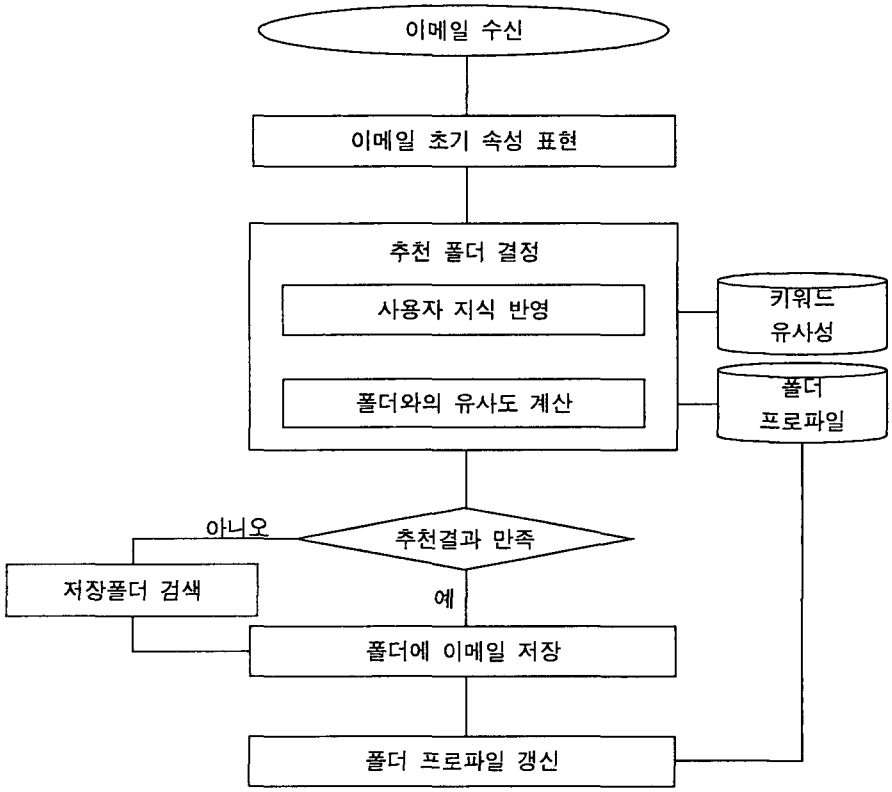
2.3 정보검색시스템의 검색 모델

정보검색시스템(Information Retrieval System)이란 사용자가 필요로 하는 정보를 수집하여 내용을 분석한 뒤 찾기 쉬운 형태로 조직하여, 정보에 대한 요구가 발생했을 때 해당 정보를 찾아 제공하는 시스템이다(Baeza-Yates & Riberrio-Neto, 1999). 정보검색시스템에서 고려되는 주요 문제는 어떤 문헌이 질의와 유사하며 어떤 문헌이 유사하지 않는가를 측정하는 것이다. 문헌과 질의의 매칭 기준은 검색 모델에 따라 다르다. 전형적인 정보검색 모델에는 문헌과 질의의 키워드의 집합으로 표현하여 집합 이론에 근거하여 유사도를 측정하는 불리언 모델(Boolean Model), 문헌과 질의를 t차원의 키워드 공간상의 벡터로 표현하여 코사인 척도로 유사도를 측정하는 벡터 공간 모델(Vector Space Model), 정보검색 문제를 확률적 틀로 해석하는 확률 검색 모델(Probability Retrieval Model)이 있다. 일반적으로 불리언 모델은 부분적인 연관/비연관을 인식하는 것이 불가능하여 가장 낮은 성능을 보이며, 가장 대중적인 검색 모델로는 벡터 공간 모델이 사용되고 있다(Baeza-Yates & Riberiro-Neto, 1999). 벡터 공간 모델은 사용자의 정보 요구와 유사한 문헌을 검색하는데 키워드의 가중치를 이용함으로써, 비교적 단순하며 검색 결과에 대한 순위화가 가능하고 벡터 수정이 용이하다는 장점이 있다(Salton, 1989). 따라서, 이메일과 폴더선호도를 TF-IDF 가중치로 표현하는 본 연구에서는 벡터 공간 모델을 사용하여 신규 메일과 기존 폴더간의 유사도를 측정한다.

3. 연구 방법론

3.1 개요

본 연구에서는 내용기반 추천(Content-based Recommendation)과 키워드 유사성(Keyword Affinity)을 이메일 환경에 맞게 결합 · 응용하여 사용자의 효율적인 이메일 관리를 지원하기 위한 방법론: KARS(Keyword Affinity based Recommender System)을 제시한다. 즉 신규 이메일에 대한 적합한 폴더를 추천함으로써 사용자가 보다 적은 시간과 노력으로 자신의 정보를 효율적으로 관리할 수 있도록 지원한다. 사용자 지



[그림 2] 이메일 폴더 추천에 관한 절차

식을 반영한 이메일 폴더 추천 프로세스는 [그림 2]과 같다.

[그림 2]에서 제시하는 폴더 추천 절차는 '이메일 초기 속성 표현", "추천폴더 결정", "폴더 프로파일 갱신"의 3 단계로 구성되며 각 단계의 기능과 적용된 기법은 다음과 같다.

Phase 1. 이메일 초기 속성 표현

신규 이메일이 도착하면 먼저 이메일의 본문 중에서 불용어를 제외한 후, 의미 있는 단어에 대한 빈도수를 계산하여 이메일에 대한 초기 단어 벡터를 생성한다.

이메일의 초기 속성은 다음의 식(1)과 같이 단어 벡터로 표현된다.

$$F(M) = (w_1, w_2, ..., w_n) \quad (1)$$

여기서, w_n 은 이메일 M 의 n번째 단어에 대한 빈도수(TF: Term Frequency)이며, 단어에 대한 가중치로 구성된 F(M)은 해당 이메일의 내용 속성을 의미한다.

Phase 2. 추천폴더 결정

추천폴더 결정 단계는 반복적인 탐색작업을 통해 신규 이메일에 가장 적합한 폴더를 결정하는 단계이다. 반복적인 탐색작업은 대상 이메일의 초

기 단어 벡터에 각 폴더의 키워드 유사성을 적용하는"사용자 지식 반영"단계와 확장된 신규 이메일의 속성과 폴더 프로파일의 속성을 비교하는 "폴더와의 유사도 계산"단계로 나누어진다.

사용자 지식 반영:

이메일 사용자는 주제별로 여러 개의 폴더를 생성하여 이메일을 관리한다. 각각의 폴더는 사용자의 지식이 반영되어 있으며 이는 키워드 유사성으로 표현된다. 즉, 사용자는 폴더의 특징을 나타내는 키워드와 키워드간의 유사성을 0과 1사이의 값으로 정의함으로써 특정 폴더를 다른 폴더와 구분 짓는다.

이렇게 정의된 키워드 유사성은 신규 이메일의 초기 벡터를 확장시킨다. 즉, 신규 메일에 포함되지 않았지만 연관성이 높은 키워드에 대한 가중치가 추가함으로써 신규 메일의 특징을 보다 정확하게 표현할 수 있다.

폴더와 유사도 계산:

신규 이메일을 관리하기에 적합한 폴더를 추천하기 위해서는 사용자가 관리하고 있는 각각의 폴더와 이메일간의 유사도를 계산해야 한다. 각 폴더의 속성 정보는 폴더 프로파일 형태로 존재하며, 폴더 프로파일은 폴더 내에 존재하는 단어들에 대한 TF-IDF가중치로 구성되어 있다. 폴더 프로파일의 생성과정은 다음과 같다.

(1) 폴더의 속한 단어의 빈도수 계산: $F(f,w)$

폴더를f, 메일을M, 메일에 속한 단어를 w라고 할 때, w에 대한 빈도수를 계산한 후 합한다.

$$F(f,w)=\sum_{M\in f}F(M,w) \tag{2}$$

(2) 폴더의 키워드 유사성 반영

폴더에 대한 사용자 지식을K, 키워드를k, k에 해당하는 유사성 가중치를a라고 할 때, 폴더의 속한 단어의 확장된 빈도수는 다음과 같다.

$$F(f,w)\leftarrow F(f,w)+K(a,k) \tag{3}$$

(3) 해당 폴더 내의 상대적인 단어 빈도수: $FF(f,w)$

단어 w가 폴더 f 내에서 차지하고 있는 가중치는 다음과 같다.

$$FF(f,w)=\frac{F(f,w)}{\sum_{w'\in F}F(f,w')} \tag{4}$$

(4) 전체 메일 대비 특정 폴더 내의 상대적인 단어 빈도수: MBED Equation.3 $TF(f,w)$

전체 메일 A에서의 단어 w의 가중치를 $FF(A,w)$라고 할 때, 단어 w가 전체 메일 A 내에서 차지하고 있는 가중치는 다음과 같다.

$$TF(f,w)=FF(f,w)/FF(A,w) \tag{5}$$

(5) 폴더의 IDF 가중치 계산

$DF(w)$는 전체 폴더 중에 특정 단어가 한번이라도 출현한 폴더의 개수를 나눈 값이며, 이 값을 기반으로 단어 w의 가중치, 즉 $IDF(w)$는

다음과 같다.

$$IDF(w) = \frac{1}{DF(w)^2} \quad (6)$$

(6) 폴더의 TF-IDF 가중치 계산

TF-IDF 가중치는 계산 식(5)와 식(6)을 통해 계산된다.

$$W(F,w) = TF(F,w) \times IDF(w) \quad (7)$$

각 폴더의 속성을 나타내는 TF-IDF가중치는 폴더 프로파일 형태로 존재하며, 신규 이메일의 속성과 폴더 속성간의 비교는 벡터 공간 모델(Vector Space Model)을 이용한다. 벡터 공간 모델에서 이메일과 폴더는 다차원의 키워드 벡터로 표현되며 이메일과 폴더의 유사도는 두 벡터의 코사인 각도를 이용하여 다음과 같이 계산된다 (Segal and Kephart, 2000).

$$SIM(M,f) = \frac{\sum_{w \in M} F(M,w)W(f,w)}{min(\sum_{w \in M} F(M,w), \sum_{w \in M} W(f,w))} \quad (8)$$

여기서, M은 신규메일이고, f 는 폴더, $F(M,w)$ 는 신규메일의 단어에 대한 빈도 벡터, $W(f,w)$는 폴더의 가중치 벡터이다. 계산된 유사도 값이 1에 가까운 폴더는 신규메일에 적합하다는 것을 의미하며, 본 연구에서는 유사도 값이 가장 높은 3개의 폴더를 추천한다.

Phase 3. 폴더 프로파일 갱신

사용자는 추천 받은 3개의 폴더 중 만족하는 폴더가 있으면 해당 폴더에 신규메일을 저장하고 그렇지 않으면 추천된 결과를 무시하고 직접 적합한 폴더를 검색한다. 폴더에 이메일이 저장되며 폴더의 프로파일은 실시간으로 갱신된다. 이러한 방식의 폴더 프로파일 갱신은 사용자의 최신의 선호도를 유지함으로써 원하는 메일에 대한 정확한 폴더를 추천 받을 수 있게 된다.

4. KARS을 이용한 폴더 추천 예제

KARS에서는 신규 이메일이 도착했을 때 사용자에게 적합한 폴더를 추천하는데 있어, 기존의 방법론: SwiftFile에 사용자 지식을 더함으로써 폴더 추천의 성능을 높이고자 하였다. 다음의 예제를 통하여 사용자 지식 즉 키워드 유사성이 신규 이메일 벡터 및 폴더의 프로파일에 어떻게 반영되는지에 대해 예를 들어 설명하고자 한다.

사용자는 이메일을 관리하기 위하여 내용에 따라 폴더별로 분류하여 저장한다. 사용자가 <표 1>과 <표 2>와 같이 2개의 폴더를 가지고 있으며, 각 폴더에는 3개, 2개의 메일이 있다. 사용자가 CRM 폴더에 "Articles"과 "Paper"간에 유사성을 0.9하였다.

신규 이메일이 수신되면, 이메일의 초기벡터는 이메일에 속한 단어의 빈도수(Term Frequency)로 <표 3>과 같이 표현된다.

사용자에게 추천할 폴더 결정하기 위해 폴더 프로파일과 신규 이메일간의 유사도를 구해야 한다. 신규 이메일 초기 벡터는 각각의 폴더 프로파일과 유사도 구하기 전에 각 폴더의 키워드 유사성을 반영해야 한다. 앞서 말한 CRM폴더의

<표 1> CRM 폴더의 이메일 벡터

폴더	이메일	CRM	Industry	Articles	Organizations	Finance	Manufacturing
CRM	1	2	1	3	1	2	0
	2	1	0	3	0	0	2
	3	2	1	2	1	0	2

<표 2> SPAM 폴더의 이메일 벡터

폴더	이메일	Finance	AD	Shopping	Present	Love
SPAM	4	2	2	2	1	1
	5	2	3	1	2	2

<표 3> 신규 이메일 초기 벡터

	Finance	Vendor	Paper	White	Case	Studies
F(신규 이메일)	1	1	2	2	1	1

"Articles"과 "Paper"간에 유사성 0.9를 신규 이메일 벡터에 반영하게 되면 신규 이메일의 초기 벡터는 <표 4>와 같이 확장된다. 초기 이메일 벡터에는 "Articles"이 없었지만 "Articles"와 "Paper" 간의 유사성 0.9를 적용하면, Articles 가중치는 Paper의 TF 가중치 2와 유사도 0.9 곱으로 구해진다.

폴더 프로파일은 폴더에 속한 메일들의 단어 벡터로 TF-IDF 가중치로 표현된다. 폴더에 속한 단어들의 TF가중치를 계산하기 위해 먼저 폴더에 속한 이메일 별로 단어 대한 TF벡터를 생성한다. 그리고 폴더 별 사용자가 정의한 키워드 유사성을 반영하여 초기 TF벡터를 확장한다. CRM 폴더의 경우, <표 1>의 CRM 폴더의 초기 이메일 벡터들은 "Articles"과 "Paper"간에 키워드 유사성 0.9를 반영하여 <표 5> 와 같이 벡터가 확장된다.

CRM 폴더의 "Articles" 이라는 단어의 TF가중치는 "Articles"이 CRM 폴더내에서 차지하는 비중을 사용자의 전체메일에서의 비중으로 나누어 구해진다. CRM 폴더 내의 각 단어의 TF값은

<표 4> CRM 폴더의 키워드 유사성을 반영한 신규 이메일 벡터

가중치	Articles	Finance	Vendor	Paper	White	Case	Studies
F(신규이메일)	1.8	1	1	2	2	1	1

<표 5> 키워드 유사성을 반영한 CRM폴더의 이메일 벡터

폴더	이메일	CRM	Industry	Articles	Organizations	Finance	Manufacturing	Paper
CRM	1	2	1	3	1	2	0	2.7
	2	1	0	3	0	0	2	2.7
	3	2	1	2	1	0	2	1.8

<표 6> CRM 폴더의 TF 가중치

Weight	CRM	Industry	Articles	Organizations	Finance	Manufacturing	Paper
$FF(CRM, w)$	0.17	0.07	0.26	0.07	0.07	0.13	0.24
$FF(A, w)$	0.10	0.04	0.17	0.04	0.12	0.08	0.15
TF	1.60	1.60	1.60	1.60	0.53	1.60	1.60

<표 7> CRM 폴더의 IDF 가중치

Weight	CRM	Industry	Articles	Organizations	Finance	Manufacturing	Paper
IDF	4	4	4	4	1	4	4

<표 8> CRM 폴더의 TF-IDF 가중치

Weight	CRM	Industry	Articles	Organizations	Finance	Manufacturing	Paper
TF	1.60	1.60	1.60	1.60	0.53	1.60	1.60
IDF	4	4	4	4	1	4	4
ation.3 $TF-IDF$	6.38	6.38	6.38	6.38	0.53	6.38	6.38

<표 6>와 같다.

CRM 폴더 내의 각 단어에 대한 IDF값은 <표 7>과 같다. "Finance"를 제외한 다른 단어들은 CRM 폴더에만 존재함으로 DF가 1/4 이고, "Finance"라는 단어는 두 개의 폴더에 다 존재함으로 DF가 1이 된다.

각 단어들의 TF와 IDF의 곱을 이용하여 CRM 폴더내의 단어 가중치, 즉 폴더 프로파일을 구하는데, 그 결과 값은 <표 8>과 같다.

마지막으로 신규로 수신된 이메일과 CRM폴더, SPAM폴더에 대한 유사성을 구한 결과는 <표 9>과 같다. 신규 이메일은 사용자의 지식을 반영하지 않은 경우에는 유사도가 0.28로 SPAM 폴더와 유사성이 와 더 가까운 것으로 나타났지만, 사용자 지식을 반영한 경우에는 유사도가 2.53으로 신규메일과 CRM 폴더의 유사성이 높게 나오는 것을 볼 수 있다. 즉, 단어의 유사성을 반영한 경우에 사용자의 요구에 맞는 추천이 이루어질 수 있음을 알 수 있다.

<표 9> CRM 폴더와 신규 이메일 간의 유사성

	SwiftFile	KARS
CRM 폴더	0.22	2.53
Spam 폴더	0.28	0.18

4. 실험

4.1 실험 데이터

이 연구의 실험 데이터는 두 명의 이메일 사용자를 대상으로 하였다. 실험에 사용된 이메일 자료 및 사용자의 폴더관리 특성은 <표 10>와 같

<표 10> 실험 대상 자료 및 폴더관리 특성

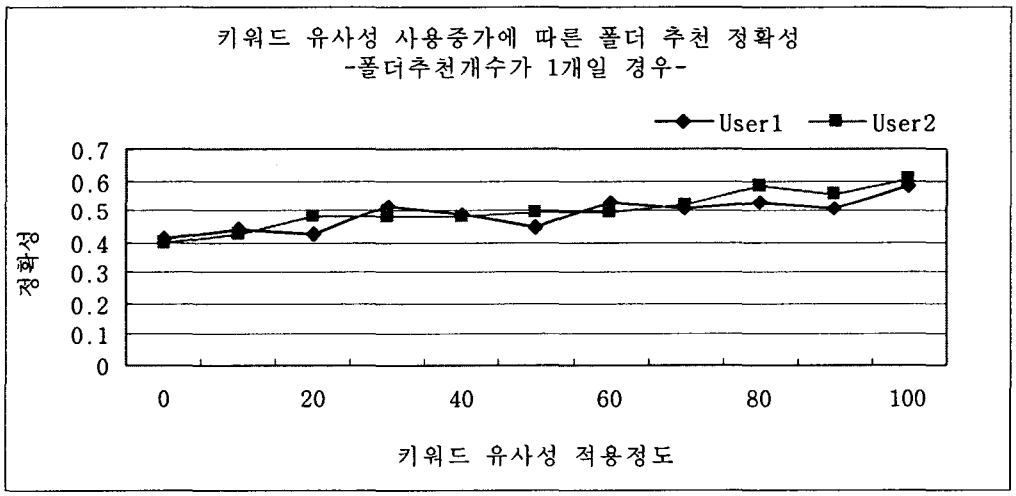
사용자	이메일 수	폴더 수	폴더별 평균 이메일 수	폴더별 평균키워드 수
사용자 1	268	16	17	6
사용자 2	318	16	20	9

다. 각 사용자의 이메일 계정 중에 하나의 계정에 속한 이메일로 사용자 1의 경우, 총 이메일의 개수는 268개이며, 사용자가 정의한 폴더의 수는 16개, 각 폴더에 포함된 이메일은 평균 약 17개 정도이다. 사용자 2의 경우, 총 이메일의 개수는 318개이며, 사용자가 정의한 폴더의 수는 16개, 각 폴더에 포함된 이메일은 평균 20개 정도이다. 실험대상에서 사용자의 메일계정에 아직 지워지지 않는 스팸 메일과 오랫동안 열어보지 않는 메일은 제외하였다.

실험을 위하여 사용자로 하여금 현재 사용하고 있는 폴더에 대한 키워드 유사성을 정의하게 하였다. 먼저, 사용자들로 하여금 신규메일이 도착했을 때 신규메일을 특정폴더에 저장하고자 할 때의 분류기준이 되는 키워드를 찾아내게 하였다. 그리고 특정의 키워드와 다른 키워드들간의 유사성을 사용자는 본인의 지식을 이용하여 0과 1사이의 다양한 비율로 표현하도록 하였다. 사용자의 1의 경우, 키워드간 유사성을 적용한 단어가 평균적으로 6개 정도이며, 사용자 2의 경우 9개 정도이다.

4.2 실험 내용

본 연구에서는 사용자 지식 즉 키워드 유사성을 폴더 프로파일 생성 시 적용하였을 때, 사용자 지식이 신규메일에 대한 폴더 추천의 정확성에 미치는 영향을 살펴보고자 하였다. 신규메일에 대



[그림 3] 추천 이메일의 개수가 1개일 경우 추천성능

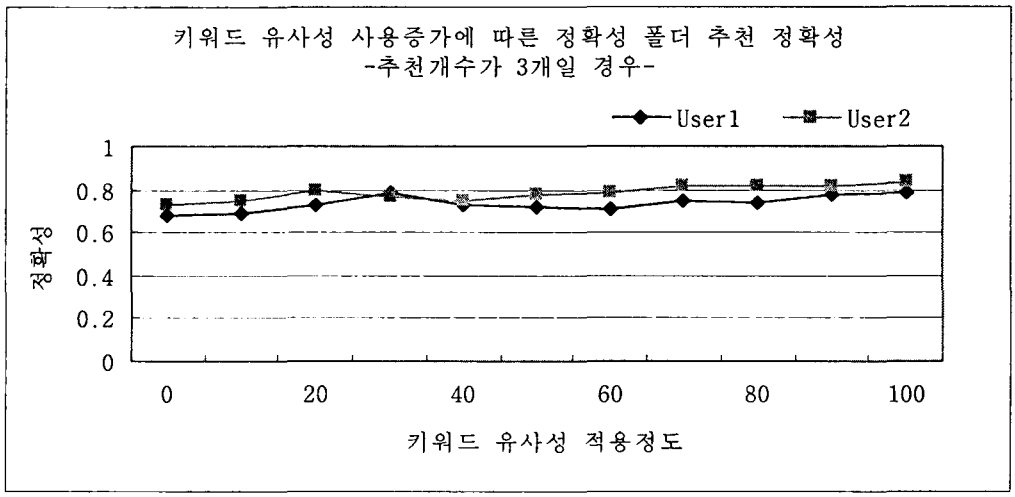
해 0~100%까지 다양한 비율로 키워드 유사성의 확대 적용해 보고, 폴더 추천의 정확성과를 파악하였다. 실험은 사용자에 추천되는 폴더의 개수 1개 인 경우, 추천폴더의 개수가 3개인 경우로 나누어 추천개수에 따라 추천의 정확성이 차이가 나는지 살펴보았다. 실험을 위한 모형 구축을 위해 데이터의 70%는 실험용으로 사용하고, 30%는 테스트용으로 사용하였다.

다음은 사용자 지식이 추천 알고리즘의 추천의 정확성에 얼마나 영향을 미치는지에 대한 실험 결과이다. 먼저 [그림 3]을 보면 유사도가 가장 높은 폴더 하나만 추천했을 경우, 사용자 지식의 반영 정도에 따라 달라지는 추천 성능을 확인해 볼 수 있다. [그림 3]의 결과를 살펴보면 사용자 지식을 사용하지 않을 경우 추천의 정확성은 0.40 정도이나, 사용자 지식을 100% 반영했을 경우 정확성은 0.60 정도로 향상 된다는 것을 알 수 있다. 또한 전반적으로 볼 때, 사용자 지식의 반영 비율이 증가할수록 정확도가 증가하는 것으로 나타나고 있다.

다음으로 이메일이 속하게 될 폴더를 3개 추천했을 경우, 사용자 지식 적용에 따른 추천 성능의 차이에 대해 살펴보았으며, 그 결과는 [그림 4]와 같다. 전체적으로 보았을 때, 사용자 지식 적용 비율이 높아짐에 따라 추천의 성능이 좋아지는 것을 확인해 볼 수 있었으며, 사용자의 지식을 사용하지 않을 경우 추천의 정확성은 0.70인 것으로 나타나고 있으나, 100% 반영한 경우 정확성은 0.81로써 0.1 정도 성능이 향상되었다.

앞서 한 개의 폴더만을 추천했을 경우와 세 개의 폴더를 추천했을 경우를 비교해 보았을 때, 사용자 지식의 반영여부에 따른 정확도의 차이는 한 개의 폴더만을 추천했을 경우 더 큰 것으로 나타나고 있다.

텍스트 기반의 내용기반 추천은 이미 축적된 선호도를 학습하여 추천 아이템을 결정한다. 선호도 정보가 충분히 축적되지 않은 환경에서 키워드 유사성을 이용하면 신규 텍스트 아이템이나 아이템에 대한 사용자의 프로파일을 보다 정확하



[그림 4] 추천 이메일의 개수가 3개일 경우 추천성능

게 표현할 수 있게 되고, 따라서 초기부터 보다 정확한 추천을 안정적으로 할 수 있다고 판단된다. 실험의 결과를 살펴보면, 이메일 환경에서 키워드 유사성은 추천을 위한 텍스트 아이템의 속성 표현 및 프로파일 확장에 대한 보조적인 수단으로 사용되어 추천성능 향상에 기여한다는 것을 볼 수 있다.

5. 결론

네트워크 기술의 발달로 인하여 사용자들은 다양하고 많은 양의 정보를 접하게 되었다. 이로 인해 사용자들은 수많은 정보 중에서 자신이 필요로 하는 정보를 찾아내기까지 많은 시간과 노력이 필요하게 되었다. 이에 본 연구에서는 대표적인 추천기법 중에 하나인 내용기반 추천과 사용자 지식에 의해 정의된 키워드 유사성을 이용하여 사용자가 보다 적은 비용으로 자신의 정보를 효율적으로 관리할 수 있도록 지원하는 방법론을 제시하였다. 특히, 동적인 환경을 가지고 있는 이메일의 분야에 적용해 봄으로써, SwiftFile방법론과 본 연구에서 제안하고 있는 KARS방법론의 비교 실험하였다. 실험 결과 SwiftFile 방법론에 비해, 본 연구를 통해 제시된 KARS 방법론의 추천 성능이 우수한 것으로 나타났으며, 특히 추천할 폴더의 수가 제한적일수록 사용자 지식을 반영했을 경우와 그렇지 않은 경우에 대한 정확성 차이가 큰 것으로 나타나고 있다. 그러나 사용자의 이메일의 양이 점점 누적이 된다면 기존의 TF-IDF 기반 프로파일이나 사용자의 지식이 부가된 프로파일이나 추천의 성능은 비슷해 질 것으로 보인다. 이메일 데이터로써 자료 수집에 따른 한계로 인해 데이터의 수가 충분하지 않아 편중된 결과가 도출되었을 수도 있다. 또한 본 연구에서 실험에 사용된 사용자 지식 즉, 폴더에 대한 사용자의 지식은 사용자가 직접 정의하는 하였으나, 일반적으로 보통의 사용자들에게 단어의 유사성 정의는 매우 번거로운 작업이며, 불필요한 정보나 과다한 정보를 정리하는 것과 마찬가지로 추가적인 시간과 노력을 요하는 작업일 것이다. 따라서 향후 연구에서는 사용자의 지식을 자동으로 생성하고 반영할 수 있는 방안에 대해 고려해 보고자 한다.

참고문헌

[1] Salton, G. and M. J. McGill, Introduction to Modern Information Retrieval, McGraw Hill Book Company, New York, 1983.

[2] Salton, G., Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley, Pennsylvania, 1989.

[3] Baeza-Yates, R. and B. Riberiro-Neto, Modern Information Retrieval, Addison Wesley Longman Inc, 1999.

[4] Balabanovic, M. and Y. Shoham, "Fab: Content-based, Collaborative Recommendation", *Communications of the ACM*, Vol. 40(1997), 66~72.

[5] Belkin, N. J. and W. B. Croft, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?", *Communications of the ACM*, Vol. 35(1992), 29~38.

[6] Basu, C., H. Hirsh, W. W. Cohen and C. Nevill-Manning, "Technical Paper Recommendation: A Study in Combining Multiple Information Sources", *Journal of*

Artificial Intelligence Research (2001), 23 1~252.

[7] Billsus, D. and M. J. Pazzani, "Learning Collaborative Information Filters", *Proceedings of Fifteenth International Conference on Machine Learning* (1998), 46~54.

[8] Chesnais P. R., M. J. Mucklo and J. A. Sheena, "The Fishwrap Personalized News System", *In IEEE Second International Workshop on Community Networking Integrating Multimedia Services to the Home*(1995).

[9] Cohen, W. W., "Learning Rules that Classify E-mail", *In Paper from the AAAI Spring Symposium on Machine Learning in Information Access*(1996).

[10] Foltz, P. W. and S. T. Dumaiz, "Personalized Information Delivery: An Analysis of Information Filtering Methods", *Communications of the ACM*, Vol. 35(1992), 51~60.

[11] Krulwich, B. and C. Burkey, "Learning User Information Interests through Extraction of Semantically Significant Phrases", *In Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*(1996).

[12] Kuflik, T. and P. Shoval, "Generation of User Profiles for Information Filtering - Research Agenda", *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*(2000).

[13] Lang, K., "Newsweeder: Learning to Filter Netnews", *In Proceedings of the 12th International Conference On Machine Learning* (1995).

[14] Lee, J. K., J. K. Kim, S. H. Kim and H. K. Park, "An Intelligent Idea Categorizer for Electronic Meeting Systems", *Group Decision and Negotiation*, Vol. 11(2002), 363~378.

[15] Pazzani, M. and D. Billsus, "Learning and Revising User Profiles: The Identification of Interesting Web Sites", *Machine Learning*, Vol 27(1997), 313~331.

[16] Pazzani, M. J., "Representation of Electronic Mail Filtering Profiles: A User Study", *Proceedings of the 5th international conference on Intelligent user interfaces*(2000).

[17] Pazzani, M. J., "A Framework for Collaborative, Content-Based and Demographic Filtering", *Proceedings of the 5th international conference on Intelligent user interfaces* (1999).

[18] Schafer, J. B., J. A. Konstan and J. Riedl, "E-Commerce Recommendation Applications", *Data Mining and Knowledge Discovery*, Vol. 5(2001), 115~153.

[19] Segal, R. B., and J. O. Kephart, "SwiftFile: An Intelligent Assistant for Organizing E-mail", *In Proceedings of the 2000 AAAI Spring Symposium on Adaptive User Interfaces* (2000).

[20] Yoon, Y. S., J. K. Lee and C. H. Han, "A Case Based e-Mail Response System for Customer Support", *Journal of Intelligent Information Systems* , Vol. 9(2003), 121~133.

Abstract

Folder Recommendation Based on User Knowledge

You, Mee* · Park, Joo Seok* · Kim, Jae Kyeong*

By the development of the network technology, the types and amount of information that users keep in contact with have been dramatically increased. As a result, users are consuming a lot of time and energy to find needed information. On this, this article presents a new methodology that can efficiently manage their information within small cost by using content-based recommendation method and keyword affinity method. By using keyword affinity method, this methodology solves the content-based recommendation method's weak point that the performance is not good within the environment that the preferences of users are rapidly changing and new contents are created continuously and the accuracy level is low until the information of preferences are sufficiently gathered.

This article carried out research on the personal e-mail environment where new information is frequently created and disappeared. Also this article assists folder recommendation for the efficient management of e-mail and verified the methodology mentioned above by an experiment to compare the performance of existing folder recommendation methods with the performance of this new method.

* School of business Administration, KyungHee University