다중 정보 여과 방법을 이용한 동적 정보 우선 순위 결정

김 진*· 윤 정섭*· 조 근식**

Dynamic Information Ranking using Multiple Information Filtering

Jin Kim*· Jeong-Seob Yoon*· Genu-Sik Jo**

요 약

인터넷의 등장으로, 끊임없이 늘어나는 정보의 양은 오히려 사용자의 정보 습득을 어렵게 만들었다. 이를 해결하기 위한 방법으로 검색된 정보에 우선 순위를 부여함으로써 사용자가 원하는 정보를 선별할 수 있는 방법이 등장하였다. 하지만, 이는 사용자의 일시적인 질의만을 가지고 정보의 우선 순위를 결정하기 때문에 사용자가 다시 판단해야 하는 부담을 안게 되었다. 이러한 문제점을 해결하기 위해, 본 논문에서는 내용 기반의 정보 검색(Content-Based Information Retrieval) 방법과 더불어 사용자의 기호를 반영하는 사용자 선호도 기반의 정보 여과(Information Filtering) 방법, 그룹 선호도 기반의 협동적 정보 여과(Collaborative Filtering) 방법을 사용하여 사용자의 요구에 보다 부합된 정보 우선 순위를 결정하는 방법을 제안한다. 제안된 방법은 사용자의 선호도 구축을 선결 조건으로 하며, 구축된 선호도는 벡터로써 표현되어 정보와의 유사도(degree of similarity) 계산에 사용된다. 제안된 방법을 실험하기 위해 MFC(Microsoft Foundation Class)관련 학습 사이트를 구현하여 사용자 등록을 받았다. 이 과정에서 사용자에게 여러 가지 프로파일을 요구하였으며, 변화하는 사용자의 기호를 반영하기 위해 지속적으로 사용자의 행동을 관찰 하였다. 이렇게 구축된 사용자 선호도를 바탕으로 제안된 방법을 실험하고 사용자의 feedback을 통해 결과에 대한 평가를 받아, 본 논문에서 제안된 방법의 타당성을 입증하였다.

Key Words : Information Retrieval, Information Filtering, Collaborative Filtering, Vector Model, Degree of Similarity, Dynamic Ranking.

---

* 인하대학교 전자계산공학과

** 인하대학교 전자계산공학과 부교수

제1 장 서 론

오늘날 인터넷 발전으로 인해 더욱더 많은 양의 정보가 일반 사용자에게 다가오고 있다. 하지만, 이러한 정보의 증가가 결코 사용자에게 정보 습득의 기회를 증가시키지는 못한다. 이러한 정보 과잉현상(information overload)을 효과적으로 완화 시켜주기 위해 정보에 우선 순위를 부여해 제공하는 방법이 등장하였다.

정보의 우선 순위를 정하는 방법으로는 사용자의 일시적인 질의와 정보와의 관계를 계산하고 그 결과를 기준으로 정보의 우선 순위를 정하는 방법이 있다. 이를 내용 기반의 정보 검색 방법(Content-Based Information Retrieval, IR)이라 한다. 하지만 이 방법은 사용자의 오래 습관이나 관심 분야에 대한 아무런 고려가 없이 때문에, 같은 질의에 대해 모든 사용자에게 획일적인 우선 순위를 결정해 줌으로써 사용자가 그 정보를 다시 판단해야 하는 부담을 안게 된다. 이러한 문제점은 사용자 선호도 기반의 정보 여과(Information Filtering, IF) 방법으로 해결되어질 수 있다. 하지만 사용자 선호도 기반의 정보 여과 방법은 새로운 정보 분야의 출연 시, 이를 효과적으로 처리하지 못하는 한계를 가지고 있다. 따라서 그 새로운 분야에 대한 다른 사용자들의 선호도를 조사하여 정보를 평가하는 협동적 정보 여과 (Collaborative Filtering, CF) 방법의 도입은 필수적이라 할 수 있다. 따라서 본 논문에서는 내용 기반의 정보 검색 방법(IR), 사용자 선호도 기반의 정보 여과 방법(IF), 그룹 선호도 기반의 협동적 정보 여과 방법(CF)을 함께 정보의 우선 순위 결정에 사용함으로써 각 방법의 한계를 극복하는 방법을 제안한다.

정보 여과 방법을 사용하기 위해서는 무엇보다도 사용자의 선호도를 구축하는 일이 선행되어야 한다. 사용자 선호도는 습득 방법과 그 특징에 따라서 분류할 수 있으며, 이는 정보 여과 시스템 구성의 또 다른 주제이기도 하다.

사용자의 요구 사항과 정보와의 관계를 계산하는 방법으로는 정보검색의 방법 중에 대표적인 벡터 모델(Vector Model)을 사용한다. 벡터 모델에서는 사용자의 질의, 사용자의 선호도, 사용자가 속한 그룹의 선호도, 그리고 정보를 벡터로써 표현하여야 한다. 이렇게 표현된 각각의 벡터를 사용하여 사용자의 요구에 부합된 정보를 검색해 내고 우선 순위를 결정한다.

본 논문의 구성을 보면, 2장에서는 정보 검색과 정보 여과의 차이점 및 상호 보안 방법과 정보 여과에서의 사용자의 선호도 구축 방법 및 그 특징을 기술 한다. 3장에서는 벡터 모델 및 정보 검색(Information Retrieval), 정보 여과(Information Filtering), 협동적 정보 여과(Collaborative Filtering)방법, 그리고 본 논문이 제안하는 이들의 상호 보안 과정을 자세히 기술한다. 4장에서는 실험을 통해 본 논문이 제안한 방법에 대한 평가를 하며, 마지막으로 5장에서 결론과 향후 연구방향을 제시한다.

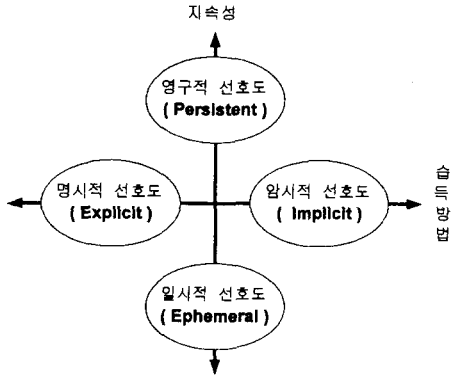
제 2 장 정보 검색과 정보 여과

사용자가 원하는 정보를 찾아주는 방법에는 크게 내용기반의 정보 검색(information retrieval) 과 사용자 선호도 기반의 정보 여과 방법(Information Filtering), 그리고 사용자 집단의 선호도를 이용한 협동적 정보 여과 (Collaborative Filtering) 방법이 있다[3]. 이들은 관심 있는 주제와 부합하는 정보를 찾아낸다는 점에서는 비슷한 역할을 한다.

2.1 정보 검색과 정보 여과의 특징

일반적으로 내용 기반의 정보 검색은 정적인 특성을 갖는 대규모의 정보 원천(static information source)을 대상으로 하다. 이러한 특징으로 인해 내용기반 정보 검색의 가장 큰 단점은 같은 질의에 대해서 모든 사용자에게 같은 정보를 획일적인 제공한다는 점이다. 정보의 획일적인 제공은 관심 분야나 취향, 직업 등 사용자 개인의 선호도는 전혀 고려되지 않았기 때문에 사용자는 다시 정보를 판단해야만 하는 부담을 가지게 된다. 이러한 내용 기반의 정보 검색의 단점을 보안하기 위해 사용자의 선호도를

함께 고려한 정보 여과 (Information Filtering) 방법이 등장하였다. 사용자 선호도 기반의 정보 여과 방법의 특징은 시간에 따라 변화하는 정보 원천(Dynamic information source)을 대상으로 한다는 점에서 내용 기반 정보 검색과는 차이점이 있다. 이러한 특징은 정보 여과 시스템이 장



[그림 2-1] 선호도의 분류

기적인 사용자의 관심을 처리하기 위해 사용자 선호도를 구축하는 것이 선행되어야 함을 의미한다[1]. 사용자의 선호도를 고려하여 정보를 제공함으로써 사용자는 정보를 다시 판단해야 하는 부담에서 벗어날 수 있다. 하지만, 정보 여과 방법은 몇 가지 한계를 가지고 있다. 사용자가 새로운 정보 분야를 요구할 경우 지능적으로 대처하지 못하는 Cold-Start 문제를 안고 있다[2]. 이는 사용자의 선호도가 충분히 구축되지 되어 있지 않기 때문에 발생한다. 그리고 사용자의 관심이 너무 한 쪽으로 치울 칠 경우 새로운 정보에 대한 접근 기회가 차단될 수 있다. 이러한 문제는 협동적 정보 여과 (Collaborative Filtering) 방법을 사용하여 해결되어 질 수 있다. 협동적 정보 여과 방법은 같은 취향을 갖는 다수의 다른 사용자의 선호도와 정보에 대한 평가를 바탕으로 이루어 진다. 따라서 사용자가 새로운 분야에 대한 정보를 요구하면, 그 정보에 대한 다른 사용자의 선호도를 바탕으로 정보를 평가할 수 있고, 때론 의외의 정보를 제공 받을 수도 있어 정보의 접근 기회를 넓혀주는 효과를 가지고 있다.

2.2 사용자 선호도의 특징

정보 검색 시스템은 정보의 내용과 사용자의 질의에 의해서 시스템이 이루어 지는 데 반해, 정보 여과 시스템은 사용자의 선호도를 습득하고 이를 학습함으로써 이루어 진다. 따라서 사용자의 선호도를 습득하는 부분은 정보 여과 시스템의 가장 중요한 구성 요소이다. 사용자의 선호도는 그 특징에 따라 몇 가지로 분류될 수 있다[4].

사용자의 선호도를 습득하는 방법에 따라 사용자에게 직접적으로 선호도를 요구하여 습득하는 명시적(Explicit) 선호도와 사용자의 행동을 관찰하여 이를 학습하여 습득하는 암시적(Implicit) 선호도로 나눌 수 있다. 또 사용자 선호도의 지속성(life-time)에 따라 일시적(Ephemeral) 선호도와 영구적(Persistent) 선호도로 나눌 수 있다. [그림 2-1]에 선호도의 특징에 따른 분류는 좌표로 표시하였다.

2.2.1 명시적(Explicit) / 암시적(Implicit) 사용자 선호도

명시적 사용자 선호도는 정보 여과 시스템이 사용자에게 직접적으로 선호도를 요구하여 그 선호도를 습득한다. 사용자에게 여러 가지 입력사항이나 선택사항을 제공하여 이를 사용자가 입력하거나, 사용자에게 정보를 제공하여 사용자가 평가(feedback)하도록 함으로써 이루어 진다. 예를 들어, 사용자의 이름, 나이, 직업, 성격, 성별 등을 입력하도록 요구할 수 있다. 하지만, 명시적 사용자 선호도는 사용자의 선호도를 얻기 위해 사용자의 직접적 노력이 필요하고 변화하는 사용자의 선호도를 반영하기 힘들다는 단점이 있다. 이런 단점에도 불구하고 명시적 선호도는 사용자의 선호도를 가장 잘 반영할 수 있다는 장점이 있기 때문에 널리 사용되어지는 선호도 습득 방법이다.

명시적 사용자 선호도와는 달리, 암시적 사용자 선호도는 정보 여과 시스템이 사용자의 행동을 관찰하고 학습함으로써 사용자의 선호도를 암시적으로 구축된다. 예를 들어, 아마존에서

는 사용자가 과거에 구매했던 책을 바탕으로 사용자의 선호도를 학습해 간다. 암시적 선호도는 사용자의 명시적인 노력 없이도 선호도를 구축할 수 있고 시간에 따라 변화하는 사용자의 선호도를 반영할 수 있다는 장점이 있다. 따라서 명시적 선호도 만큼 정확히 사용자의 선호도를 표현할 수는 없지만, 명시적 선호도가 갖는 단점을 보완하는 방법으로 사용되어 진다.

2.2.2 일시적(Ephemeral) / 영구적(Persistent) 사용자 선호도

일시적 사용자 선호도는 단지 현재의 정보 여과 시에만 고려되어 사용되는 선호도이다. 이는 순간순간 변화하는 사용자의 선호도를 반영할 수 있으며, 과거의 상태가 현재의 상태에 영향을 주지 않으므로 사용자의 선호도 왜곡을 막을 수 있다. 하지만, 일시적 사용자 선호도는 짧은 기간동안 사용되어 짐으로 암시적 방법에 의해서 구축하기는 매우 어렵다.

이에 반해 영구적 선호도는 사용자의 과거의 선호도가 현재의 상태까지 영향을 미칠 수 있다. 이는 사용자의 오랜 습관이나 변화하지 않는 관심을 반영하는 데 효과적으로 사용된다.

이렇게 구축된 사용자 선호도는 사용자의 질의와 더불어 사용자 요구사항을 표현하는 중요한 데이터이다. 사용자 선호도 기반의 정보 여과에서는 이를 기반으로 정보와의 유사도를 계산하며, 그룹 선호도 기반의 정보 여과에서는 사용자 선호도를 바탕으로 사용자 그룹을 형성하여 정보와의 유사도를 계산하다. 따라서 사용자 선호도 구성은 사용자의 선호도를 최대로 반영할 수 있도록 체계적이며 논리적으로 구성되어야 한다.

제 3 장 우선 순위 결정을 위한 모델링

사용자에게 우선 순위가 부여된 정보를 제공하기 위해서는 사용자의 요구사항과 정보와의 유사도(degree of similarity)를 측정하여야 한다. 본 논문에서는 사용자의 요구사항을 사용자의 질의, 사용자의 선호도, 사용자 그룹의 선호도로 구분하여 구성하였다. 이들 사용자의 요구사항은 서로 독립적으로 존재하기 때문에 각각 별도의 벡터로서 모델링하여 정보와의 유사도를 측정할 수 있다. 따라서 사용자의 요구 사항과 정보와의 유사도를 측정하는 방법으로 정보검색의 대표적인 방법 중 하나인 벡터 모델을 사용한다.

3.1 벡터 모델(Vector Model)

벡터 모델이란 사용자 요구사항과 정보를 index term의 집합으로 표현한 후 각 index term에 가중치(weight)를 부여 함으로써 이루어진다. 부여된 weight를 바탕으로 사용자 요구사항과 정보와의 유사도를 측정하여 그 정보가 사용자가 원하는 정보와 얼마나 부합되는 지를 판단한다. 또한 이 유사도는 각 정보의 순위(ranking)를 결정하는 데 사용되어 진다. 유사도를 계산하기 위해서는 사용자의 요구 사항과 정보를 벡터로써 표현 하여야 한다.

사용자의 요구 사항을 벡터로 표현한 식은 [식 3-1]과 같다.

q : 사용자요구사항(or query)

$w_{i,q}$: q의 i번째 index term 에 부여된 weight

q를index term vector 를사용하여표현하면

$$q = (w_{1,q}, w_{2,q}, \ldots w_{t,q}) \qquad [식\ 3-1]$$

정보를 벡터로 표현한 식은 [식 3-2]와 같다.

d_j : j번째 정보 (or document)

t : index term 의 수

k_i : i 번째 index term

$K = \{k_1, \ldots k_t\}$: 모든 index term 의 집합

$w_{i,j}$: d_j에서 k_i의 weight

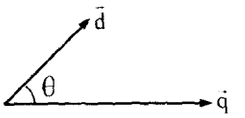
d_j를 index term vector 를 사용하여 표현하면

$$\vec{q}_j = (w_{1,j}, w_{2,j}, \ldots w_{t,j}) \qquad [식\ 3-2]$$

사용자의 요구사항과 벡터와 정보벡터와의 유사도를 구하는 공식은 [식 3-3]과 같은 코사인 계수공식을 사용한다. 이를 [그림 3-1]과 같이 표현할 수 있다.

$$sim(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j \bullet \vec{q}}{|\vec{d}_j| \times |\vec{q}|} \qquad [식3-3]$$

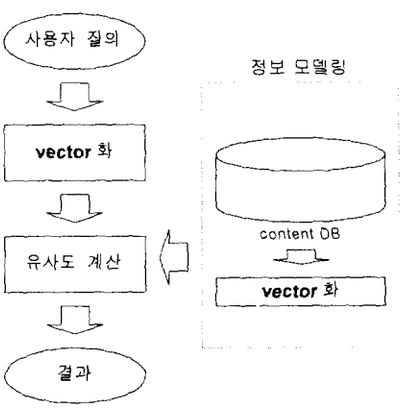
$$= \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w^2_{i,j} \times \sum_{j=1}^{t} w^2_{i,q}}}$$



[그림 3-1] $sim(d_j, q)$ 은 두 벡터 $\vec{d}j$ 와 $\vec{q}$ 의 cosine θ 값으로 표현된다.

3.2 정보 검색(Content-based Information Retrieval) 모델

내용기반의 정보 검색은 사용자의 질의와 정보와의 관계를 계산하여 정보를 검색하는 방법이다. 내용기반의 정보 검색은 사용자가 원하는 정보를 찾기 위해서 사용자의 요구가 포함된 일시적인 질의에 초점을 맞추고 있다. 사용자의 질의를 분석하여 벡터화하고 다시 정보 벡터와의 유사도를 계산함으로써 사용자의 요구에 가장 부합된 정보를 검색해 낸다. 사용자의 질의와 정보와의 유사도를 구하는 과정은 [그림 3-2]와 같다.



[그림 3-2] 질의에 의한 정보 검색 과정

사용자 질의와 정보를 벡터화 하는 방법은 3.1절에 나와있다. 질의 벡터와 정보 벡터와의 유사도를 구하기 위해서는 각 벡터의 index term에 weight를 부여해 주어야 한다.

정보의 index term에 weight를 주는 방법으로는 TF-IDF방식을 사용하여 계산한다. TF는 정보의 크기에 따라 그 값이 변할 수 있으므로 정규화 하여 사용한다. TF와 IDF는 구하는 방법은 [식 3-4]와 [식 3-5] 에 나와 있다.

TF : Term Freqeuncy

$freq_{i,j}$: d_j에 나타난 k_i의 빈도수

$\max_i freq_{i,j}$:

d_j에 나타난 index term 중 가장 빈도수 많은 index term 의 빈도수

$$f_{i,j} = \frac{freq_{i,j}}{\max_i freq_{i,j}} : 정규화된\ TF \qquad [식 3-4]$$

IDF : Inverse document frequency

N : 정보의 총 수

n_i : k_i가 나타난 정보 (문서)의 수

$$idf_i = \log \frac{N}{n_i} \qquad [식 3-5]$$

정보의 index term에 weight를 주는 방법은 [식 3-6]과 같다.

$$w_{i,j} = f_{i,j} \times \log \frac{N}{n_i} \qquad [식3-6]$$

사용자의 요구사항에 index term을 주는 방법은 Salton and Buckely가 제안한 방법으로 [식 3-7]과 같다[5].

사용자의 요구 : q

$freq_{i,q}$: 정보에 대한 q안의 term k_i의 발생 빈도수

$\max_i freq_{i,q}$: q에서 나타난 term중 가장 빈도수가 많은 term의 빈도수

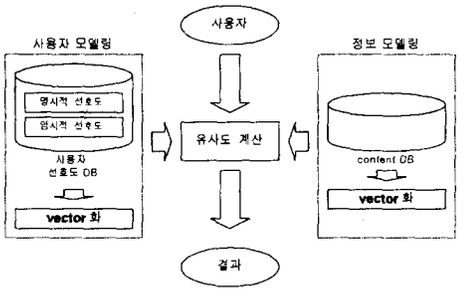
$$w_{i,q} = \left(0.5 + \frac{0.5\ freq_{i,q}}{\max i\ freq_{i,q}}\right) \times \log \frac{N}{n_i} \qquad [식3-7]$$

3.3 정보 여과 (Information Filtering) 모델

정보 여과 시스템은 다음과 같은 과정으로 이루어 진다.

사용자의 관심도를 파악하여 사용자 프로

파일을 구축하는 사용자 모델링 단계와 정보를 표현하기 위한 정보 특징 추출 단계가 우선 수행된다. 다음으로 검색된 정보가 사용자의 관심 정보인지를 판단하기 위한 사용자 선호도와 문서와의 유사도를 측정하며, 피드백을 통한 학습으로 사용자의 특성에 프로파일을 일치시켜 변화하는 사용자의 관심을 적용해 간다.



[그림 3-3]

사용자 선호도에 의한 정보 여과 과정

사용자 선호도를 벡터로 표현하기 전에 앞서 정보의 특징을 추출하여 n가지의 category로 정보를 분류한다. [식 3-8]은 category집합을 나타내고 있다.

n : category 의 수

$$C = \{c_0, c_1, \ldots, c_k c_{n-1}\} \qquad [식 3-8]$$

각 정보는 하나 이상의 category에 포함되어진다. 이는 곧 정보가 category를 사용하여 벡터화될 수 있음을 의미한다. [식 3-9]은 정보

n : category 의 수

$$\vec{d} = \{c_0, c_1, \ldots, c_k, \ldots, c_{n-1}\} \qquad [식 3-9]$$

를 category를 사용하여 벡터화한 식이다.

i번째 사용자의 선호도 벡터($\vec{p}_i$)는 사용자에게 선호도를 직접적으로 요구하여 구축하는 명시적(Explicit) 선호도 벡터($\vec{p}_{E,i}$)와 사용자의 행동을 관찰하고 학습하여 구축하는 암시적(Implicit) 선호도 벡터($\vec{p}_{I,i}$)로 구성된다. 사용자 선호도의 각 index term은 정보의 분류 기준인 category와 일치하도록 구성한다. 이는 사용자 선호도 벡터와 정보 벡터와의 유사도를 구하기 위함이다.

명시적(Explicit) 선호도 벡터($\vec{p}_{E,i}$)의 index term이 가질 수 있는 값은 boolean 값을 갖는다고 가정한다. 이는 사용자의 선호도를 표현하는 데 무리가 있지만, 사용자가 정확히 자신의 선호도를 표현하기 힘들 뿐만 아니라, 표현이 가능할 지라도 자칫 사용자에게 잘못된 선호도 기입을 초래할 수 있기 때문이다. i번째 사용자의 명시적 선호도 벡터는 [식 3-10]같이 표현된다.

n : category의 수

v_k : 명시적 선호도 벡터의 요소(0 or 1)

$$\vec{p}_{E,i} = (v_0, v_1, \ldots v_k, \ldots, v_{n-1}) \qquad [식 3-10]$$

암시적 사용자 선호도 벡터($\vec{p}_{I,i}$)는 사용자의 행동을 관찰하여 구축되어진다. 특히 사용자가 해당 정보에서 의미 있는 행동(예를 들어, 물건을 구입할 때)을 할 경우, 그 정보가 속한 category에 관심이 있는 것으로 간주 하고 사용자의 선호도에 반영하게 된다. i번째 사용자의 암시적 선호도 벡터는 [식 3-11]같이 표현된다.

n : category 의 수

w_k : 암시적 선호도 벡터의 요소($0 < w_k$)

$$\vec{p}_{I,i} = (w_0, w_1, \ldots w_k, \ldots, w_{n-1}) \qquad [식 3-11]$$

하지만, 어느 특정 category에 대한 정보가 너무 많거나 적을 경우, 사용자의 암시적 선호도는 category의 구성에 따라 왜곡될 수 있다. 따라서 이를 정규화하기 위해 각 category 마다 다른 weight값을 주어 이를 사용자 선호도 변화에 반영해 주어야 한다. k번째 category의 weight값을 구하는 공식은 [식 3-12]에 나타나 있다.

$Sum(c_k)$: k번째 category의 문서수

N : 정보의 총수

$$weight(c_k) = \frac{1}{\sqrt{Sum(c_k) \bullet N}} \qquad [식 3-12]$$

또한 사용자의 선호도는 시간의 흐름에 따라 변화하므로 현재의 사용자 행동은 과거의 사용자 행동 보다 더욱 비중있게 사용자의 선호도에 반영되어져야 한다. 따라서 사용자가 category c_k에 속하는 정보에 대한 사용자의 암시적 선호도 벡터의 index term (w_k)의 변화는 [식 3-13]에 의해서 이루어 진다.

α : 상수
β : 시간에 따른 감쇄계수 ($0 < \beta < 1$)
t : 현재의 날짜 (date)
t-1 : w_k가 마지막으로 생신된 날짜
diff : t와 t-1의 날짜 간격

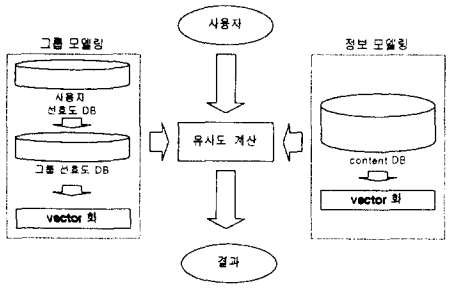
$$w_{k,t} = \text{weight}(c_k) \bullet \alpha + \beta^{diff} \bullet w_{k,t-1} \quad [식\ 3-13]$$

[식 3-10]와 [식 3-11]에 의해서 표현된 명시적 사용자 선호도 벡터와 암시적 사용자 선호도 벡터를 사용하여 i번째 사용자 선호도 벡터를 표현하면 [식 3-14]와 같다.

$$\vec{p}_i = \vec{p}_{E,i} + \vec{p}_{I,i} \quad [식\ 3-14]$$

3.4 협동적 정보 여과 (Collaborative Filtering) 모델

협동적 정보 여과는 정보의 내용을 직접 분석할 필요 없이 사용자들의 관계만을 이용하여 정보를 추천하며, 정보의 추천 범위를 넓혀 뜻하지 않은 정보를 추천할 수도 있다. 또한 정보의 내용 뿐만 아니라 정보의 우수성에 따라 정보를 추천할 수 있다.



[그림 3-4]

그룹 선호도에 의한 정보 여과 과정

그룹 선호도 벡터를 구축하기 위해서는 사용자를 그룹화하는 과정이 필요하다. 본 논문에서는 사용자를 그룹화하는 방법으로 Supervised Learning Vector Quantization 방법을 사용한다 [6]. 이 방법은 미리 여러 개의 그룹을 구축한 후, 새로운 사용자 들어 올 경우 그 사용자의 선호도 벡터와 가장 부합되는 그룹에 포함 시키는 방법이다. 한 사용자는 동시에 여러 개의 그룹에 포함 될 수 있다. 본 논문에서 사용하는 그룹은 정보의 분류 기준인 category에 근거하여 구축 하였다. 사용자 그룹의 집합은 [식 3-15]과 같다.

n : category 수

$$G = \{g_0, g_1, \ldots, g_k, \ldots, g_{n-1}\} \quad [식\ 3-15]$$

어느 특정 정보에 대한 그룹의 선호도를 측정하기 위해, 각 정보는 자신을 방문했던 사용자 그룹의 행동을 관찰하고 학습해야 한다. 이렇게 학습된 결과는 각 정보와 사용자 그룹간의 선호도를 측정하는 데 사용되어 진다. j번째 정보가 가지고 있는 각 사용자 그룹과의 선호도를 나타내는 index 벡터($\vec{g}$)는 [식 3-16]과 같다. 그리고 각 index term에 weight를 부여하는 식은 [식 3-17]에 나와 있다.

벡터의 각 index term은 그룹에 속한 사용자가 해당 정보에서 의미 있는 행동을 했을 때 갱신 된다. 즉, index 벡터 $\vec{g}$는 해당 정보와 각 그룹(g_k)과의 유사도를 의미한다.

n : category 수

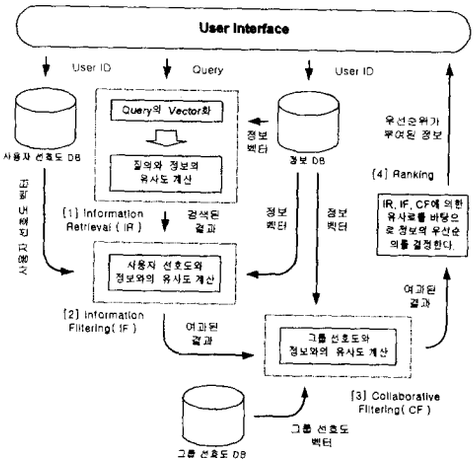
$$\vec{g}_j = (u_{j,0}, u_{j,1}, \ldots u_{j,k}, \ldots, u_{j,n-1}) \quad [식\ 3-16]$$

$$\text{weight}(u_{j,k}) = \frac{u_{j,k}}{\sum u_k} \quad [식\ 3-17]$$

3.5 다중 선호도 벡터를 사용한 우선 순위 결정(Ranking)

정보의 우선 순위 결정은 각 모델에서 제시된 사용자 벡터(질의, 사용자 선호도, 그룹 선호도)와 정보 벡터와의 유사도 측정을 통해서 이루어 진다. 내용 기반의 정보 검색 모델에서는 사용자의 질의와 정보를 단어(word)위주의 index term으로 표현하고 각각에 weight를 부여함으로써 유사도를 측정한다. 사용자 선호도 기반의 정보 여과 모델에서는 사용자의 선호도와 정보를 category 위주의 index term으로 표현하여 유사도를 측정한다. 그리고 그룹 선호도 기반의 정보 여과 모델에서는 각 정보에 그룹에 대한 index 벡터를 두어 그룹과 정보와의 유사도를 측정하다.

[그림 3-5]는 본 논문에서 제안하는 정보 검색 시스템의 구조를 보여준다.



[그림 3-5] 전체 시스템 구조

본 시스템의 정보 우선 순위 결정 과정은 크게 네 단계로 이루어 진다.

i) 첫번째 단계: 사용자의 질의를 입력 받아 그 질의를 벡터로 표현(식3-1)하여 정보 벡터(식3-2)와의 유사도(식3-3)를 계산한다(내용기반의 정보 검색, Information Retrieval).

ii) 두번째 단계: 사용자의 신분(User ID)을 파악하여 사용자 선호도 벡터를 추출(식3-14)한다. 이렇게 추출된 사용자 선호도 벡터와 첫번째 단계에서 검색된 결과의 정보 벡터(식3-9) 사이의 유사도를 계산하다(사용자 선호도 기반의 정보 여과, Information Filtering).

iii) 세번째 단계: 사용자가 속한 그룹 선호도 벡터(식3-16)를 추출하여 첫번째 단계에서 검색된 정보와의 유사도를 계산하다(협동적 정보 여과, Collaborative Filtering).

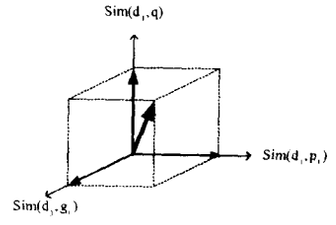
iv) 마지막 단계: 각 단계에서 계산된 유사도를 종합하여 정보의 우선 순위(Ranking)를 결정하고 사용자에게 제공한다.

각 단계에서의 유사도는 서로 전혀 다른 방법에 의해서 계산되었으므로, 각 유사도를 독립된 벡터로써 표현할 수 있다. 따라서 유사도의 합은 [식 3-18]과 같다.

$Sim(d_j, q)$: j번째 정보와 질의(query)의 유사도
$Sim(d_j, p_i)$: j번째 정보와 i번째 사용자의 유사도
$Sim(d_j, g_i)$: j번째 정보와 i번째 그룹의 유사도

$$Sum(Sim(d_j, q), Sim(d_j, p_i), Sim(d_j, g_i)) = \left| Sim(d_j, q) + Sim(d_j, p_i) + Sim(d_j, g_i) \right| = \sqrt{Sim(d_j, q)^2 + Sim(d_j, p_i)^2 + Sim(d_j, g_i)^2} \quad [식\ 3-18]$$

이러한 관계를 그림으로 표현하면 [그림 3-6]과 같다.

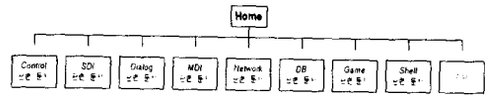


[그림 3-6]
$Sim(d_j, q)$, $Sim(d_j, p_i)$, $Sim(d_j, g_i)$ 합

제 4 장 실험 및 평가

4.1. 실험 환경

본 논문에 대한 실험을 하기 위해 MFC(Microsoft Foundation Class)관련 학습 웹 사이트를 구축하였다[9]. MS NT4.0 환경을 기반으로 MS-IIS 5.0을 Web Server로, MS-SQL 7.0을 DB Server로, VBScript를 프로그램 언어로 선택하였다. 4개월간 웹 사이트를 운영하며 3600여명의 사용자를 등록 받았으며, 1000여 page의 html 문서를 구축하였다. 웹 사이트의 구성은 9개의 category별로 분류된 계층적 구조를 사용하였다.



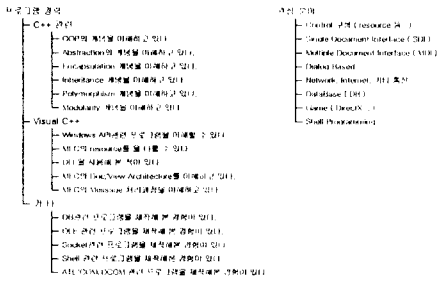
[그림 4-1]웹 사이트의 구성도

4.2 사용자 선호도 구축

정보 여과에 필수적인 사용자의 선호도를 구축하기 위해 사용자에게 다양한 입력 사항을 요구 하였다. 일반적인 사용자 프로파일(이름, 나이, 직업 따위)과 전문적 사용자 프로파일(프

로그램 경험 및 관심분야)로 나누어서 입력하도록 하였다. 전문적 사용자 프로파일은 사용자가 쉽게 자신의 선호도를 표현할 수 있도록 체크박스를 사용하였다.

전문적 사용자 프로파일에 대한 세부 사항은 [그림 4-2]에 자세히 나와있다.



[그림 4-2] 사용자 프로파일 세부 항목

4.3. 평가 기준

본 시스템의 성능을 평가하는 기준으로 정확도(Precision)와 재현율(Recall), 그리고 F-measure를 이용한다[5]. 정확도는 시스템에 의해서 검색된 결과 중에 사용자가 원하는 결과와 얼마나 일치하는 지를 측정하다. 정확도는 사용자의 선호도가 얼마나 잘 학습 되었는 지를 나타내는 척도이다. Recall은 사용자가 예상한 결과와 실제로 시스템에 의해서 검색된 결과와 얼마나 일치하는 지를 측정한다. F-measure는 정확도와 Recall의 가중치 조합으로 나타내면 그 범위는 0과 1 사이의 값을 가진다.

$$F\text{-}Measure = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \qquad [식4-1]$$

4.4. 결과 및 평가

본 논문에서 제안된 시스템과 기존의 검색 방법과의 성능을 비교하기 위해 사용자에게 세 가지 경우로 나누어 검색 결과에 대한 feedback을 받았다. feedback은 검색된 결과 중 사용자가 원하는 정보가 1~5순위에 있으면 정확도를 높였으면, 6순위 밖에 있으면 재현율을 낮추었다.

경우-I) 내용기반의 검색(Information Retrieval, IR) 방법만 사용했을 경우.

경우-II) 내용기반(Information Retrieval, IR)방법과 정보 여과(Information Filtering, IF)방법을 함께 사용하였을 경우.

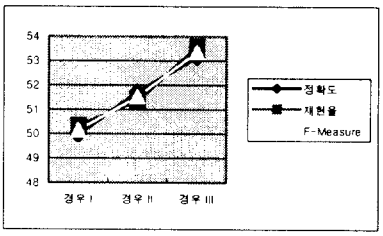
경우-III) 내용기반(Information Retrieval, IR) 방법과 정보 여과(Information Filtering, IF) 방법 그리고 협동적 정보 여과(Collaborative Filtering, CF) 방법을 함께 사용하였을 경우.

실험에 대한 신뢰도를 높이기 위해 최근 5일간 5회 이상 방문한 사용자만을 실험에 1회에 한하여 참여 시켰다. 총 567명을 실험에 참여하였다. 실험에 따른 정확도(Precision), 재현율(Recall), 그리고 F-Measure의 측정치는 [표 4-1]에 나타나 있다

	경우 I	경우 II	경우 III
정확도	49.97%	51.72%	53.14%
재현율	50.34%	51.35%	53.57%
F-Measure	50.15%	51.53%	53.35%

[표 4-1] 실험 결과

실험 결과에서 경우-III의 검색 결과가 경우-I 보다는 정확도는 3.17%, 재현율은 3.23%, F-Measure는 3.20% 앞서 있으며, 경우-II보다는 정확도는 1.42%, 재현율은 2.23%, F-Measure는 1.82% 앞섰다.

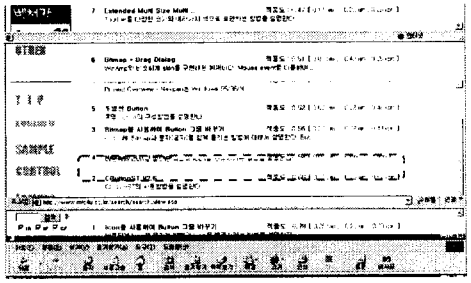


[그림 4-3] 각 경우에 대한 평가 비교

따라서 협동적 정보 여과(CF) 방법이 내용기반의 정보 검색(IR)의 단점과 사용자 선호도 기반의 정보 여과(IF) 방법의 한계를 효과적으로 보안할 수 있음을 입증하는 결과이다.

또한 제안된 방법을 사용함으로써 의외의 결과가 주어 지기도 한다. 이는 일반적으로 검색해 내기 힘든 정보를 비슷한 선호도를 가진 다

른 사용자들의 영향으로 인해 검색해 내기 때문이다. [그림 4-4]는 협동적 정보 여과 방법에 의해서 낮은 순위의 정보가 높은 우선 순위의 정보로 바뀐 결과를 보여 주고 있다.



[그림 4-4] 질의어 "button"에 대한 정보 우선 순위 변화

제 5 장 결론 및 향후 연구

본 논문에서는 정보의 과잉현상(Information Overload)을 효과적으로 완화하기 위한 방법으로 다중 정보 여과(Multiple Information Filtering) 방법을 사용한 동적 정보 우선 순위 결정(Dynamic Information Ranking) 방법을 제안하였다. 이 방법은 내용 기반의 정보 검색 방법뿐만 아니라, 사용자 선호도 기반의 정보 여과(Information Filtering) 방법과 그룹 선호도 기반의 협동적 정보 여과(Collaborative Filtering) 방법을 함께 사용하여 보다 사용자 요구 사항에 부합된 정보를 제공하는 방법이다.

사용자의 선호도가 정보 우선 순위 결정에 반영됨으로써, 같은 질의에 대해서도 각 사용자는 자신의 취향에 맞는 우선 순위를 제공받을 수 있게 되었으며, 다시 여기에 사용자 그룹의 선호도를 추가 시켜 사용자 선호도 기반의 정보 여과 방법이 가지는 여러 가지 단점들을 보안하였다. 이러한 일련의 과정을 수행하기 위해 사용자의 질의와 선호도, 그리고 사용자 그룹의 선호도를 벡터로 모델링하여 정보와의 유사도를 계산하는 방법을 제시하였다.

향후 연구로서, 사용자의 선호도 구축방법에 대한 보다 향상된 관찰 방법 및 학습 방법의 연구가 필요하다. 증가하는 정보의 양 만큼이나 사용자의 기호도 급속히 변하고 있다. 따라서 사용자의 선호도를 알아내는 다양한 방법들이 연구 되어져야 하고 이를 바탕으로 더욱 향상된 정보 우선 순위가 제공되어져야 한다.

참고 문헌

[1] 송미란, 김교정, "사용자 그룹을 이용한 효과적인 정보 여과 및 학습 방법에 관한 연구", 한국 정보 과학회, p63 ~ p65, 1999.

[2] Yezdi Lashkari, Max Mertal, Pattie Maes, "Collaborative Interface Agent", Readings In Agents, Morgan Kaufmann Publishers, p111~p116, 1998.

[3] Nathaniel Good, J.Ben Schafer, Joseph A. Konstan, Al Borchers, Badrul Sarwar, JonHerlocker, and John Riedl, "Combining Collaborative Filtering with Personal Agents for Better Recommendations", AAAI , 1999.

[4] Schafer, J.B.Konstan, Riedl, "Recommender Systems in E-Commerce." Proceedings of the ACM Conference on Electronic Commerce, 1999.

[5] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, "Modern Information Retrieval", Addison-Wesley Publish, p27~p30, 1999.

[6] Sergios theodoridis, Konstantinos Koutroumbas, "Pattern Recognition", Academic Press, 1998.

[7] Marko Balabanovic and Yoav Shoham , "Content-based, Collaborative Recommendation", Communications of ACM 40(3), p66~p72, 1997.

[8] http://www.mfc4u.co.kr/