퍼지 및 지능시스템학회 논문지 2005, Vol. 15, No. 4, pp. 443-450

웹 사용자의 선호도 추출을 위한 지능모델 설계 및 평가

Design & Evaluation of an Intelligent Model for Extracting the Web User' Preference

김광남* · 윤희병* · 김화수**

KwangNam Kim*, Heebyung Yoon* and Hwa-Soo Kim**

* 국방대학교 전산정보학과

** 아주대학교 정보통신대학원

요 약

본 논문에서는 웹 사용자의 선호도를 추출하기 위한 지능적 모델을 제안하고 이에 대한 평가결과를 제시한다. 이를 위해 현재 정보검색엔진의 문제점을 분석하고, 선호도 가중치를 학습기에 반영한다. 이것은 키워드에 의한 단어별 빈도수에 의존하지 않고 지능적으로 사용자의 행동유형을 학습하게 함으로써 질의에 대한 결과집합을 사용자의 의도에 맞게 제공하는 메커니즘이다. 다음으로 선호도 유행성에 대한 개념과 고려요소를 제안하며, 선호도 추출 알고리즘과 이에 대한 예를 제시한다. 또한 행동유형 추출을 위한 지능모델을 설계하고 HTML 색인과 선호도 결정 지능학습과정을 제안한다. 마지막으로 선호도를 적용한 후의 문서 랭킹 측정결과를 비교함으로써 본 논문에서 제안한 모델의 타당성을 검증한다.

Abstract

In this paper, we propose an intelligent model for extraction of the web user's preference and present the results of evaluation. For this purpose, we analyze shortcomings of current information retrieval engine being used and reflect preference weights on learner. As it doesn't depend on frequency of each word but intelligently learns patterns of user behavior, the mechanism provides the appropriate set of results about user's questions. Then, we propose the concept of preference trend and its considerations and present an algorithm for extracting preference with examples. Also, we design an intelligent model for extraction of behavior patterns and propose HTML index and process of intelligent learning for preference decision. Finally, we validate the proposed model by comparing estimated results(after applying the preference) of document ranking measurement.

Key words : Preference Weight, Information Retrieval Engine, Intelligent Model, Behavior Pattern

1. 서 론

정보검색엔진은 사용자로 하여금 필요한 정보를 효율적으로 검색할 수 있도록 도와주는 가장 일반적인 도구이다. 일반적인 정보검색엔진은 사용자의 질의를 받으면 적게는 수천에서 많게는 수억 개의 관련 페이지를 사용자에게 보여준다. 이들 중에 실제로 사용자가 원하는 페이지는 몇몇에 불과하며, 대부분의 페이지들은 검색도 하지 않는 상태에서 버려지곤 한다. 또한 랭킹 알고리즘은 하나의 질의에 대해서 모든 사용자에게 동일한 페이지를 서열화하여 내림차순으로 보여줌으로써 불필요한 시간적 낭비를 초래하곤 한다.

이러한 서비스는 사용자들의 요구사항을 만족시키지 못한다. 그 이유는 일반적인 정보검색엔진이 다음과 같은 문제점들을 포함하고 있기 때문이다[1]. 첫째, 너무 많은 페이지가 검색결과로 나타난다는 것이다. 일반적으로 사용자들은 단지 첫 페이지나 또는 몇 개의 페이지만을 확인한다. 관련 결과값으로 나타난 무수히 많은 페이지들이 사용자에게 보여 지지만 그들 중에 실제로 사용자가 원하는 문서는 얼마 되지 않으며, 또한 얼마 되지 않는 관련문서를 찾기 위해 많은 페이지들을 다 확인하기 위해 많은 시간을 허비하지 않는다. 둘째, 질적 수준이 낮은 페이지들이 많이 검색된다는 것이다. 대용량의 결과집합에 있는 적은 정보나 링크를 포함하는 짧은 페이지들이 대표적인 경우라 할 수 있다. 셋째, 중복되는 페이지가 존재한다는 것으로 이것은 비슷하거나 똑같은 페이지들이 결과집합에 반복해서 나타난다는 것이다. 넷째, 다양한 의미를 갖는 질의어에 대해서는 관련 없는 문서들도 결과집합으로 나타난다는 것이다. 다섯째, 스팸에 노출되어 있어 텍스트 분석 기술만을 채용한 검색엔진에 대해서는 치명적인 영향을 미친다는 점이다.

본 논문에서는 위에서 제시한 문제점들 중에 첫 번째와 두 번째 문제를 선호도 유행성을 기반으로 사용자의 행동유형을 분석하고, 그런 다음 지능적으로 반복학습을 통해 관련 문서를 결과집합으로 표현하는 모델을 제시한다. 이를 위해 선호도 유행성에 대한 개념과 적용방안을 제시하며 수식과 예를 통하여 제안된 모델을 구체화한다. 다음으로 행동유형 추출을 위한 지능모델의 설계를 통하여 일반적인 정보검색엔진에 적용하기 위한 방법을 제시하고 선호도를 적용하기 전

접수일자 : 2004년 8월 17일
완료일자 : 2005년 6월 15일

과 적용하고 난 후의 데이터 분석을 통하여 결과를 비교 검토한다. 또한 선호도 추출 알고리즘과 선호도를 결정하기 위한 지능형 학습과정을 제시하며 제안한 모델과 알고리즘을 바탕으로 측정한 문서간의 검색결과를 비교 검토한다. 마지막으로 결론 및 향후 연구방향을 제시한다.

2. 관련연구

사용자는 다양한 행동패턴을 가지고 있다[2]. 이러한 행동유형을 분석하기 위해 대표적으로 사용하는 것이 웹로그 마이닝이다[3][4][5]. 웹로그 마이닝은 세 가지의 주요 단계인 전처리, 지식추출, 결과분석을 포함한다. 이것은 웹 사이트에 접근했던 사용자의 정확한 목록을 결정하고 이들 사용자의 세션을 분석하여 재구성하기 위한 것이다. 데이터를 전처리하기 위해 데이터 융합, 데이터 정제, 데이터 구조화, 데이터 요약화 등의 과정을 거친다[5]. 데이터 융합 과정은 로그파일을 합하는 과정이며, 데이터 정제 과정은 불필요한 데이터를 제거하는 과정이다. 데이터 구조화 과정은 사용자, 세션 등에 의해 비구조화되어 있는 요청을 그룹화하는 과정이며, 요약화 과정은 파일을 관계형 데이터베이스에 전송하여 요청 수준별로 데이터 일반화하고 사용자 세션에 대해 계산된 데이터를 종합하는 과정이다. 지식추출 과정은 연관분석이나 순차적 패턴 등을 이용하여 중복되지 않은 데이터에 대한 연관지식을 발견하는 과정이다[4]. 결과분석은 구조화된 자료를 그룹화하여 시각적으로 표현하는 단계로서 drill- down, roll-up, slice 그리고 dice 등의 기법이 활용된다[3].

Srivastava는 서버접근 로그, 참조 로그, 사용자 등록 또는 프로파일 정보와 같은 다양한 데이터 소스 통합을 포함하여 서버측의 데이터 컬렉션에 대한 분석결과를 제시하였다[6]. 이러한 정보는 수집된 데이터로부터 사용자의 유일한 키를 식별하는 어려운 문제점을 해결해 주며, 사용자 데이터, 사용자 행동모델로부터 사용자의 세션이나 트랜잭션을 식별할 수 있는 중요한 요인을 발견해 준다. Cooley는 정보와 유형 발견으로부터 웹로그 분석 처리절차를 자세히 제시하였다[7]. 데이터 정제를 통하여 불필요한 로그 데이터 예를 들면, URL 내에 연결된 파일이름 중 GIF, JPEG, JPG와 같은 그림파일을 갖는 엔트리들을 제거하고, 프락시 서버나 로컬 캐시에 의해서 잘못 인식되어 질 수 있는 부분은 쿠키, 캐시 버스팅 그리고 사용자 등록정보를 통하여 해결한다. 다음으로 경로분석, 연관규칙, 순차적 패턴, 클러스터링 그리고 분류 등의 방법을 적용하여 규칙을 추출한 다음 정보를 제공해 준다.

또한 웹로그에 유사도 분석을 적용하여 개인화에 관한 정보를 추출하는 기법도 제공되고 있다[8][9]. WebPersonalizer는 웹서버 로그와 하이퍼텍스트를 이용하여 개인화 시스템 에이전트를 벡터 공간상에서 코사인 유사도 분석을 통한 결과를 보여 준다[8]. 특히, 동적인 프로파일을 이용한 개인화를 위해서 유사한 속성을 갖는 ID를 그룹으로 설정하여 별도로 해당 그룹 ID를 관리하는 방법도 제공한다[9]. 동적인 사용자 프로파일을 생성하기 위해 사용자 정보에 관한 분류뿐만 아니라 각 사용자 세션에 대해서 자주 접근했던 문서들의 통계데이터를 추출하여 사용자의 행동유형에 대한 변화에 대처할 수 있는 메커니즘도 제공한다.

Velásquez는 사용자의 선호도를 분석하고 반영하기 위해 웹콘텐츠 마이닝과 웹사용자 마이닝을 결합한 새로운 방법을 제시하였다[10]. 사용자 세션을 분석하여 어떤 페이지를 방문하고 더 선호하는지를 방문자 선호도 분석을 통해서 비교된 측정값을 구하고, 그런 다음 근접 행위를 갖는 사용자 세션 그룹을 찾기 위해 클러스터링 알고리즘을 이용하여 분류하고 방문할 사용자의 선호도에 대한 예측방법을 제시하였다.

3. 선호도 유행성 개념 및 고려요소

3.1 개념

본 논문에서 정의한 선호도 유행성이란 사회적으로 유행하거나, 개개인이 선호하는 것은 다른 사람들도 또한 선호하여 모방한다는 것을 의미하며, 이것은 어떠한 정보를 얻고 싶지만 찾기 위한 방법을 모를 때 경험적인 지식을 갖고 있는 사람에게 자문을 구하여 정보를 얻는 것과 유사한 의미를 갖는다. 즉, 정보검색에서도 마찬가지로 원하는 정보를 얻고자 할 경우, 정보를 제공해 주는 정확한 주소를 모를 때 다른 사람들이 선호하는 정보를 제공해 줌으로써 사용자가 효율적으로 정보를 얻을 수 있도록 하는 것이다. 이러한 선호도 유행성을 반영하기 위해서 본 논문에서도 웹사용자 마이닝을 웹콘텐츠 마이닝에 적용하여 설계한다.

3.1 선호도 유행성 고려요소

웹사용자 마이닝을 수행할 때 해결하기 어려운 문제점 몇 가지가 존재한다[6]. 첫째, 단일 IP 주소와 다중 서버 세션으로서 ISP는 웹을 접근하기 위한 통로인 프록시 풀을 가지고 있어, 하나의 웹 사이트를 접근하는 여러 사용자를 가질 수 있다는 것이다. 둘째; 다중 IP와 단일 서버 세션 부분에서 어떤 ISP나 개인 도구들은 임의적으로 한 사용자에게 몇 개의 IP 주소들 중에 하나를 이용하여 각각의 요청을 할당하는데, 이 경우 단일 서버 세션은 다중 IP 주소를 가질 수 있다는 것이다. 셋째, 다중 IP 주소와 단일 사용자의 경우에 다른 머신에서 웹을 접근하는 사용자는 세션에서 세션으로 다른 IP 주소를 가질 수 있다는 것이다. 이것은 동일 사용자로부터 반복 방문하는 것에 대한 추적을 어렵게 만든다. 마지막으로, 다중 에이전트와 단일 사용자의 경우로서 하나 이상의 브라우저를 사용하는 사용자는 비록 같은 웹 브라우저를 사용하더라도 다중 사용자로 나타나는 문제점을 가지고 있다. 이러한 경우에는 웹사용자 마이닝에 대한 분석이 어렵기 때문에 향후 지속적으로 연구되어야 할 분야이며, 이 4가지 문제점에 대한 해결방안은 본 논문의 연구대상에 제외한다.

3.3 웹로그 분석시 고려사항

사용자가 원하는 문서를 정확히 분석하기 위해서는 웹사이트 체류기간에 대한 임계값을 정하는 것이 가장 중요하다. 임계값은 기본적으로 실험 결과값인 30분으로 규정하고 있지만[11], 본 논문에서는 선호도가 0인 사이트 중에 최대로 체류한 시간을 임계값으로 고려한다. 이에 대한 알고리즘이 그림 1에 나타나 있다.

4. 선호도 추출 알고리즘

4.1 선호도 및 상관계수

처음에 문서가 색인되는 경우는 선호도 가중치가 반영되지 않은 상태이므로 각 문서에 대해 용어빈도수에 의한 가중치만을 적용한다.

```
let t be surfing time on the web
let threshold value be  boundary of timeout
threshold value = 0
select max(time) where time = surfing time not
        related to preference in web log
threshold value = max(time)
if(t >= threshold value)
   regarded as user-wanted site
else
   regarded as wrong site
```

그림 1. 체류기간 결정 알고리즘
Fig. 1. Decision-making algorithm for timeout

그리고 사용자의 질의 및 웹사용자 마이닝을 통하여 학습 과정을 거친 후 개인 선호도에 대한 가중치를 랭킹 알고리즘에 반영한다. 개인 선호도 가중치를 얻어내기 위해 다음과 같은 관계를 정의한다.

$$A_{m,n} = \begin{pmatrix} count_{1,1} & \cdots & count_{1,n} \\ \cdot & \cdot & \cdot \\ count_{m,1} & \cdots & count_{m,n} \end{pmatrix}$$ 이라 할 때,

A는 m x n 행렬이 되며, 여기서 m은 사용자가 사용한 질의어 수이고 n은 모든 질의어에 대한 결과집합 D로 추출된 문서의 수라고 가정한다. q_i $(1 \leqq i \leqq m)$는 i번째 질의어이며, $d_j$$(1 \leqq j \leqq n)$는 j번째 문서라 한다. $count_{i,j}$ $(1 \leqq i \leqq n,\ 1 \leqq j \leqq m)$는 i번째 질의어로 나타난 문서의 결과집합 중에서 j 번째 문서를 사용자가 선택한 횟수라 한다. 만약 q_i 질의로 나타난 결과집합으로부터 d_j 문서를 한번도 선택하지 않았다면 $count_{i,j}=0$이 된다. 질의어 q(q_i에 대한 생략 표현)에 대해 문서 d_j가 선택될 확률인 $\Pr(d_{q,j}|q)$는 다음과 같이 표현된다.

$$\Pr(d_{q,j}|q) = \frac{|d_{q,j}|}{\sum_{j=1}^{n}|d_{q,j}|} = \frac{1}{n} \quad (1)$$

여기서 $|d_{q,j}|$는 질의 q에 대한 결과문서 중 1개를 의미하므로 1의 값을 가지며, 식 (1)의 분모는 1부터 n개까지이므로 n의 값을 가진다. 그리고 이 n값은 추출된 전체 문서의 수인 $|D|$와 같다고 할 수 있다.

또한, $Pref_{q,j}$를 $d_{q,j}$에 대한 선호도(사용자의 기댓값)라고 하면 식 (1)에 의해 다음과 같이 정의된다.

$$Pref_{q,j} = \frac{|d_{q,j}| \times count_{q,j}}{n} = count_{q,j} \ / \ |D| \quad (2)$$

왜냐하면 $|d_{q,j}|$의 수는 1이고 n은 $|D|$이기 때문이다. 따라서 질의에 대한 문서 선호도는 최종적으로 다음과 같이 정의된다.

$$Pref(D_j|q) = Count_{q,j}/|D_q| \quad (3)$$

다음은 학습 전 랭킹결과 집합과 학습 후 랭킹결과 집합의 변화량인 ∇W에 대해 살펴본다. R_{before}를 학습 전 랭킹결과 집합, R_{after}를 학습 후 랭킹결과 집합이라 놓으면 랭킹 변화량 ∇W는 다음과 같다.

$$\nabla W = R_{before} - R_{after} \quad (4)$$

여기서 $\nabla W=0$이면 변화량이 더 이상 없으므로 학습을 종료하고, $\nabla W\neq 0$이면 재학습을 실시하며 학습 전의 선호도와 학습 후의 선호도 연관성을 두 변수간 상호연관성을 측정하는데 가장 많이 사용되는 Pearson 상관계수를 이용하여 계산한다. Pearson 상관계수를 구하는 식은 다음과 같다.

$$\Gamma = \frac{\sum(X-\mu_x)(Y-\mu_y)}{N\sigma_x\sigma_y} \quad (5)$$

식 (5)를 이용하여 학습 전 선호도와 학습 후 선호도에 대한 상관계수를 γ_{pref}라고 놓으면 다음과 같이 γ_{pref}을 나타낼 수 있다.

$$\gamma_{pref} = \frac{\sum(pref_{q,j}^{before} - \overline{pref}_q^{before})(pref_{q,j}^{after} - \overline{pref}_q^{after})}{N\partial_{before}\partial_{after}} \quad (6)$$

여기서 $pref_{q,j}^{before}$는 학습 전의 질의 q에 대한 i번째 문서 선호도를 나타내며 $pref_{q,j}^{after}$는 학습 후의 문서 선호도를 나타낸다. N은 총 문서수이며 ∂_{before}는 학습 전의 선호도에 대한 표준편차를 나타내고, ∂_{after}는 학습 후의 선호도 표준편차를 나타낸다. $\overline{pref}_q^{before}$는 학습 전의 질의 q에 대한 전체 문서 선호도의 평균이며 $\overline{pref}_q^{after}$는 학습 후의 전체 문서 선호도의 평균을 나타낸다.

4.2 선호도 추출 예

위에서 설명한 선호도 추출 알고리즘을 예를 들어 설명하면 다음과 같다. 각 질의에 대한 문서의 선호도를 분석하기 위해 마우스 클릭-스트림을 통해 추출해낸 $q \times d$의 빈도수 매트릭스 A를 다음과 같이 가정한다.

$$A = \begin{pmatrix} 2 & 1 & 3 & 0 \\ 1 & 3 & 1 & 0 \\ 0 & 2 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

위의 매트릭스 A는 식 (3)을 이용하여 매트릭스 $Pref_A$로 다음과 같이 표현될 수 있다.

$$Pref_A = \begin{pmatrix} 0.5 & 0.25 & 0.75 & 0 \\ 0.25 & 0.75 & 0.25 & 0 \\ 0 & 0.5 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0 & 0 \end{pmatrix}$$

$Pref_A$를 반영 후 사용자의 선호도를 다시 분석한 결과를 $Pref_A'$ 이라고 하며 다음과 같이 가정한다.

$$Pref_A' = \begin{pmatrix} 0.3 & 0.4 & 0.5 & 0.1 \\ 0.25 & 0.75 & 0.3 & 0 \\ 0 & 0.4 & 0.5 & 0.3 \\ 0.4 & 0.25 & 0 & 0.2 \end{pmatrix}$$

$Pref_A$와 $Pref_A'$간 랭킹 변화량이 발생하므로 질의별 문서의 상관계수인 γ_{pref}를 구하면 다음과 같다.

$$\gamma_{pref} = \begin{pmatrix} 0.63 \\ 0.74 \\ 0.57 \\ 0.63 \end{pmatrix}$$

그림 2와 그림 3은 하나의 문서를 대상으로 유사도 분석을 이용하여 가중치를 적용한 결과와 선호도를 적용한 결과를 색인용어에 대한 가중치로 비교하여 나타낸 것이다. 50개의 질의에 대해 선호도를 적용한 경우와 유사도 분석을 이용한 경우의 문서랭킹을 비교한 결과, 일부는 그림 2에서처럼 랭킹 순위값에 변화가 없는 경우도 있고, 그림 3에서처럼 랭킹 순위값에 변화가 나타난 경우도 있다.

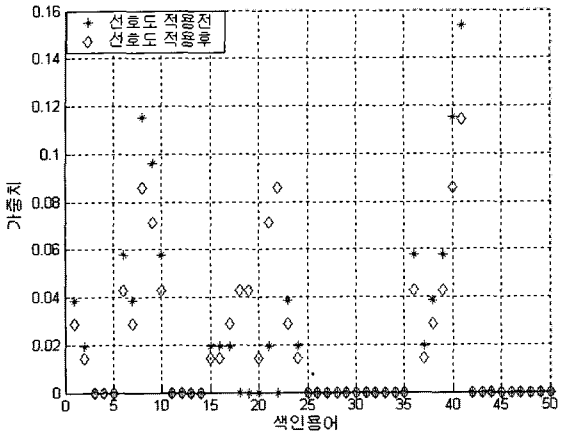

그림 2. 선호도 적용 후 랭킹순위에 변화가 없는 경우
Fig. 2. The case that ranking did not change after applying the preference

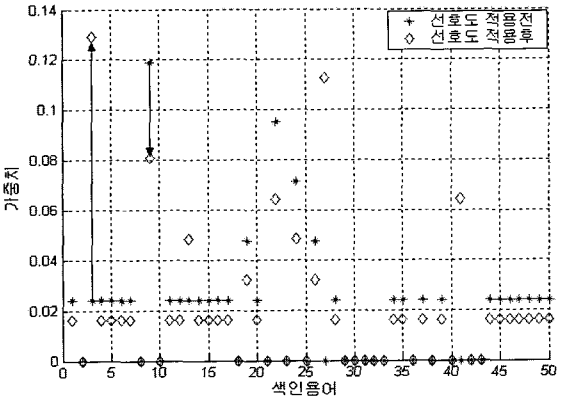

그림 3. 선호도 적용 후 랭킹순위에 변화가 있는 경우
Fig. 3. The case that raking changed after applying the preference

그림 3에 선호도 적용 전과 선호도 적용 후의 색인용어의 가중치 순위 변화가 나타나 있다. 이 그림에서 9번째 용어가 선호도 적용 전에는 가중치 값이 가장 높으므로 이 용어를 이용하여 질의를 할 경우 관련 문서가 나타날 가능성이 가장 높다. 그러나 선호도를 적용한 후에는 그림 3에서 보듯이 3번째 용어의 가중치 값이 가장 높게 나타난다는 것을 알 수 있다. 이 경우 우리는 3번째 색인용어를 이용하면 질의에 가장 관련이 많은 문서를 제공받을 수 있다는 것을 알 수 있다. 이처럼 선호도 가중치를 적용한 결과는 유사도 분석에 의한 가중치 결과값과 차이를 가질 뿐만 아니라, 사용자가 선호하는 문서에 가중치를 더 부여하여 결과집합에 랭킹을 부여하므로 검색결과의 정확성 면에서 더 우수하다는 것을 알 수 있다. 이러한 결과는 일반적으로 사용자가 선호하는 문서에 대해서는 다른 사용자들도 더 선호한다는 선호도 유행성을 잘 반영해 주고 있는 결과라고 볼 수 있다.

4.3 선호도 추출 알고리즘

위와 같이 저장된 데이터는 학습을 통하여 각 문서의 용어 가중치를 다시 적용하는데 이용된다. 그림 3에서 보는 바와 같이 선호도 결합 프로세스로부터 유사도 분석과 선호도 값을 결합하여 재 계산된 값을 랭킹에 반영하여 사용자 인터페이스를 통하여 그 결과를 보여준다. 그림 4는 선호도를 추출하기 위한 학습과정을 보여준다.

```
Let s be a t-by-d term-document frequency matrix
Let p be a t-by-d term-document preference matrix
Let ∇ return boolean result of ranking change
p=0;
while(∇=not true)
  for selected document di
    for(j=1;j≤t;j++)
      extract counti,j from web-log by user through query
      where termj=query
      p(i,j)=counti,j
    end for
  end for
  s=s+p
  analyse-similarity(s)
end while
store final p to repository
```

그림 4. 선호도 추출 알고리즘
Fig. 4. Algorithm for extraction of preference

선호도 추출 알고리즘을 위해 s를 용어와 문서간의 유사도 분석을 한 벡터값으로 정의하고, p를 용어와 문서 간에 웹로그 분석을 통한 선호도 벡터값으로 정의한다. 그리고 ∇은 선호도를 반영한 후 웹로그 분석을 통하여 지속적인 학습과정을 거친 후 랭킹에 변화가 있는지를 확인하는 부울값으로 정의한다. 그러면 최초 웹로봇에 의해 다른 사이트로부터 가져온 문서는 선호도 값이 반영되지 않으므로 선호도 매트릭스 p를 0으로 초기화한다. 사용자의 마우스 클릭-스트림을 이용하여 저장된 웹로그를 분석하여 선호도 매트릭스 p에 해당하는 값을 할당하고, 변경된 p의 값을 s에 할당한다. 변경된 s값을 이용하여 용어와 문서의 유사도 분석을 통하여 가중치를 다시 계산한다. 만약 랭킹에 변화가 있으면 다시 반복하고, 랭킹에 변화가 더 이상 발생하지 않을 때까지 이와 같은 학습과정을 계속한다.

5. 선호도 추출 지능모델 설계

5.1 선호도 추출 프레임워크

일반적인 정보검색엔진에서 추가적으로 선호도를 구현하기 위해 선호도 유행성을 적용한 지능모델의 구성이 그림 5에 나타나 있다. 사용자 질의에 대한 응답 결과를 인터페이스를 통해 사용자에게 제시하면 사용자는 결과집합으로 나타난 문서 중 관심있는 문서를 먼저 선택하고 그런 다음 해당 웹 사이트로 이동하여 원하는 정보를 얻을 것이다.

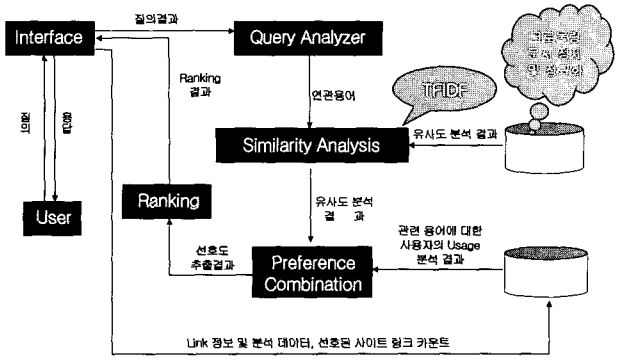

그림 5. 선호도 유행성을 적용한 지능모델
Fig. 5. Intelligent model based on preference trend

이때 마우스 클릭-스트림을 웹로그에 저장하며 질의용어와 URL 간의 관계를 이용하여 선호도를 추출한다. 웹로그의 분석은 정제과정을 거쳐 웹사용자 마이닝을 통해 색인된 데이터의 가중치 값에 반영된다.

지금까지 설명한 전체적인 시스템의 프로세스별 인터페이스가 UML 시퀀스 다이어그램을 이용하여 그림 6에 도시되어 있다. 프로세스별로 상호연관이 되어 있으며, 질의를 이용하여 얻어진 데이터를 질의 분석으로부터 랭킹까지 순차적으로 나타나 있다. 그림 6의 과정은 그림 4에서 설명한 내용과 같이 랭킹에 변화가 있으면 반복해서 수행되는 프로세스를 표현한 것이다.

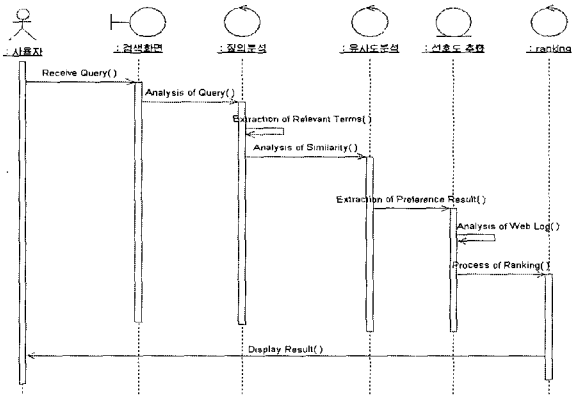

그림 6. 시스템 프로세스 간 인터페이스
Fig. 6. Interface between system processes

5.2 Web Log 정제

개인선호도에 가중치를 부여하기 위해서는 먼저 Srivastava[6]가 제시한 방법을 이용하여 웹사용자 마이닝을 수행한다. 이를 위해서 먼저 웹로그에 대한 분석을 수행하고 이를 통해 새로운 정보를 얻어내는 방법이 필요하다. 먼저 웹로그를 정제하고 필터링을 수행한 후 필요한 데이터를 다시 가공한다. 그런 다음 순차적 패턴과 연관규칙 등을 이용하여 웹페이지의 경로를 식별하고 사용자가 선호하는 페이지를 식별하기 위한 작업을 수행해야 한다.

웹로그를 정제하는 과정이 그림 7에 도시되어 있다. 즉, 실(raw)데이터에서 불필요한 부분을 제거하여 정제된 데이터를 저장하고, 다시 정제된 데이터를 통합한 다음 연관규칙, 순차적 패턴, 클러스터링, 분류 등을 이용하여 별도의 자료를 추출한다.

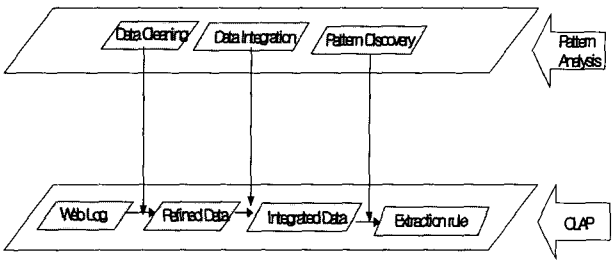

그림 7. Web Log 정제과정
Fig. 7. Refinement process of web log

웹로그는 그림 7에 도시한 바와 같이, 정제된 로그를 OLAP와 같은 도구를 이용하여 결과집합을 저장소에 저장한다. OLAP와 같은 도구는 분석 지향적이며 주제 지향적이고 복잡한 질의 지향적이면서 다차원 데이터 모델을 사용하는데 매우 효과적이다[12]. 그 전에 전처리 과정을 거치는 데 정제된 데이터를 만들기 위해서 불필요한 로그 자료는 제거된다. 그림 8은 사용자의 행동유형을 기록한 IIS의 W3SVC 웹로그 자료의 일부를 보여주고 있다.

```
#Fields: date time c-ip cs-username s-ip s-port cs-method cs-uri-stem cs-uri-query sc-status
2002-09-03 02:07:16 8.112.4.218 - 8.112.4.218 80 GET /Board/login.aspx - 200 Mozilla/4.0+(co
2002-09-03 02:07:16 8.112.4.218 - 8.112.4.218 80 GET /Board/MyBoard.css - 200 Mozilla/4.0+(c
2002-09-03 02:07:40 8.112.4.218 - 8.112.4.218 80 POST /Board/login.aspx - 302 Mozilla/4.0+(c
2002-09-03 02:07:40 8.112.4.218 - 8.112.4.218 80 GET /Board/MemberBoard/insert.aspx - 200 Mo
2002-09-03 02:07:40 8.112.4.218 - 8.112.4.218 80 GET /Board/MemberBoard/MemberBoard.css - 20
2002-09-03 02:07:40 8.112.4.218 - 8.112.4.218 80 GET /aspnet_client/system_web/1_1_4322/WebU
2002-09-03 02:07:40 8.112.4.218 - 8.112.4.218 80 GET /Board/images/new.gif - 200 Mozilla/4.0
2002-09-03 02:07:40 8.112.4.218 - 8.112.4.218 80 GET /Board/images/list.gif - 200 Mozilla/4.
2002-09-03 02:07:45 8.112.4.218 - 8.112.4.218 80 GET /Board/list.aspx - 200 Mozilla/4.0+(com
2002-09-03 02:07:45 8.112.4.218 - 8.112.4.218 80 GET /Board/images/lastest.gif - 200 Mozilla
2002-09-03 02:07:52 8.112.4.218 - 8.112.4.218 80 GET /Board/insert.aspx - 200 Mozilla/4.0+(c
2002-09-03 02:09:31 8.112.4.218 - 8.112.4.218 80 POST /Board/insert.aspx - 200 Mozilla/4.0+(
2002-09-03 02:09:49 8.112.4.218 - 8.112.4.218 80 POST /Board/insert.aspx - 200 Mozilla/4.0+(
2002-09-03 02:09:59 8.112.4.218 - 8.112.4.218 80 POST /Board/insert.aspx - 200 Mozilla/4.0+(
```

그림 8. IIS의 W3SVC 웹로그 자료
Fig. 8. Web log data of W3SVC in IIS

로그 자료의 일부 문서는 불필요한 데이터인 그림파일, 스크립트, 데이터 로딩시 필요한 dll 등의 내용들도 포함하여 기록된다. 이러한 자료들은 사용자의 행동유형을 분석하는데 도움을 주지 못하는 자료이므로 정제과정에서 제거된다. 정제된 데이터는 순차적 패턴을 이용하여 사용자의 웹사이트 경로를 추적할 수 있으며, 또한 질의 내용에 대해서 가장 많이 방문한 웹사이트의 통계를 분석하여 선호도 가중치를 만드는데 도움을 준다.

5.3 HTML 색인

색인과정은 a, an, the와 같은 색인에 불필요한 불용어를 제거한 용어와 이와 관련된 URL을 추출한다. 결과로 나타난 색인용어와 URL 값은 웹로그 정제시 관련 URL을 비교하여 학습과정에서 반영된다. XML 파싱에 대한 결과값을 저장하는 알고리즘은 최단거리와 깊이탐색을 통한 색인 알고리즘을 이용하여 이미 증명되었다[13]. 그림 9에 이와 같은 html 문서를 파싱하여 저장하는 화면을 보여주고 있다.

5.4 질의 분석

질의 분석기는 사용자의 질의어를 분석하여 연관된 용어를 분석하고 해석하는 과정으로서, 색인된 용어와 문서의 연관관계를 추출하기 위한 전 과정으로서 이용된다. 유사도 분

석기는 색인된 데이터를 유사도 분석 기법을 이용하여 문서와 용어 간의 관계를 식 (1)에 의해서 계산된 결과값을 적용한다.

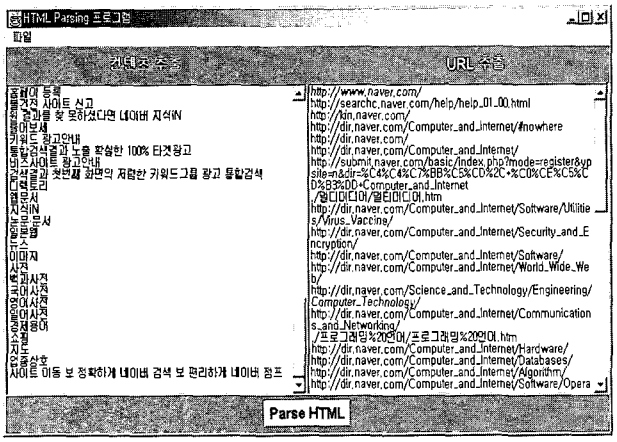

그림 9. HTML 색인
Fig. 9. HTML indexing

5.5 선호도 결합

선호도 결정을 위한 지능적인 학습과정이 UML의 활동 다이어그램을 이용하여 그림 10에 도시되어 있다. 즉, 시작점으로부터 질의에 대하여 유사도 분석을 한 후 선호도 값을 계산하고, 가중치에 변화가 있는지 확인한다. 만약 가중치에 변화가 있다면 다시 유사도 분석과정을 통하여 가중치를 다시 계산하고, 가중치에 변화가 없으면 학습을 종료한다.

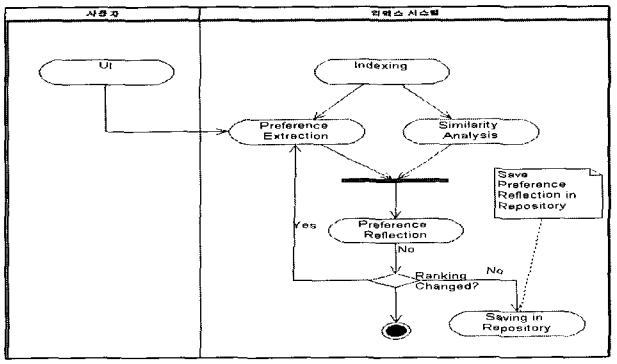

그림 10. 선호도 결정을 위한 지능적 학습과정
Fig. 10. Intelligent learning process for deciding the preference

이와 같이, 선호도 변화량의 계산을 통해 학습과정을 반복할 것인지 아니면 종료할 것인지 자동으로 결정하는 지능적인 학습과정을 이용하여 사용자의 행동유형을 분석하고 이를 반영하며 필요시 다시 계산할 수 있는 방법을 제공해 줌으로써 사용자가 요구하는 정보에 보다 근접한 결과를 제공해 줄 수 있다.

6. 평 가

선호도 추출을 위해 50개의 웹문서를 대상으로 일반 유사도 분석과 사용자 선호도 반영 후의 결과값을 비교 하였다. 그림 11과 그림 12에 사용자 질의에 대한 선호도 반영 전과 반영 후의 랭킹 결과값이 50개의 문서에 대해 분포로서 표시되어 있다.

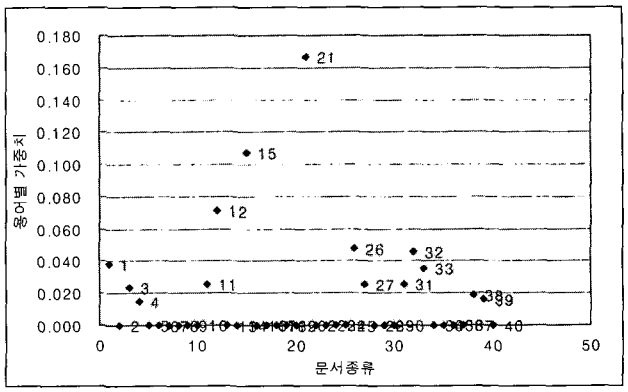

그림 11. 선호도 가중치 반영 전 문서 분포
Fig. 11. Distribution of documents before reflection of preference

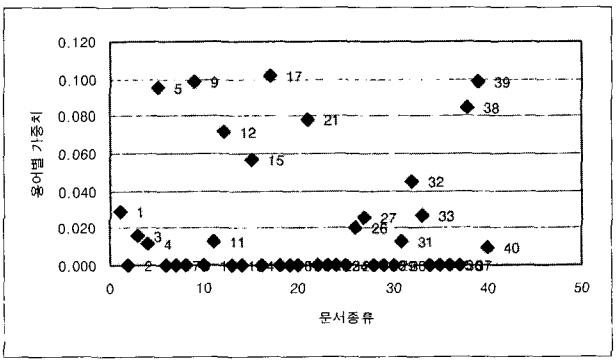

그림 12. 선호도 가중치 반영 후 문서 분포
Fig. 12. Distribution of documents after reflection of preference

위 그림에서도 알 수 있듯이 선호도 반영 전에는 21번 문서가 질의에 대해 가장 큰 가중치를 가지므로 이 문서가 가장 우선적으로 검색결과로 나타나고 15번, 12번의 문서 순으로 나타난다. 그러나 선호도를 반영한 후에는 17번 문서가 가장 큰 가중치 값을 가지므로 먼저 검색결과로 나타나고 50번, 9번, 6번의 문서 순으로 나타난다는 것을 볼 수 있다. 이처럼 사용자의 행동유형을 지능적으로 학습하고 추출함으로써 사용자가 선호하는 문서를 제공할 수 있다.

7. 결론 및 향후 연구방향

본 논문에서는 웹사용자의 선호도 추출을 위한 지능적인 모델과 이에 대한 평가결과를 제시하였다. 이를 위해 일반적인 정보검색엔진에 더하여 유사도 분석모듈과 선호도 결합모듈, 웹로그를 선호도에 반영하기 위한 모델 등을 제안하였다. 지능모델의 타당성 검증을 위해 실 데이터와의 비교 결과를 제시하였으며, 선호도 추출 알고리즘과 선호도 결정을 위한 지능적 학습과정을 또한 제안하였다. 선호도 가중치를 이용하여 문서와 용어 간 가중치를 유사도 분석 데이터와 비교한

결과 사용자 선호도와 유사도 분석에 의한 가중치에는 많은 차이가 있음을 실험적 데이터 분석을 통해서 검증하였다.

결론적으로 본 논문에서 제안한 사용자 행동유형 분석 기반의 선호도 반영은 연속적인 학습과정을 통하여 사용자가 요구하는 정보를 제공해 줄 수 있는 지능적 모델임을 입증하였다. 향후 세션기반의 유사도 측정의 단점을 극복하기 위한 개인화 적용방법과 선호도 유행성 기반의 정보발견 기법이 요구된다.

참 고 문 헌

[1] Feng Guozhen, Cheng Xueqi, Bai Shuo, "SAInSE : An Intelligent Search Engine Based on WWW Structure Analysis," *15th International Parallel and Distributed Processing Symposium*, pp.1734-1740, 2001.

[2] Catherine Stones and Stephen Sobol, "DMASC : A Tool For Visualizing User Paths THrough A Web Site," *13th International Workshop on Database and Expert Systems Applications*, pp.389-393, 2002.

[3] Osmar R. Zaïane, Man Xin and Jiawei Han, "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs," *Research and Technology Advances in Digital Libraries, IEEE International Forum*, pp.19-29, 1998.

[4] Yew-Kwong Woon, Wee-Keong Ng, Xiang Li and Web-Feng Lu, "Efficient Web Log Mining for Product Development," *Internaltional Conference on Cyberworlds*, pp.294-301, 2003.

[5] Doru Tanasa and Brigitte Trousse, "Advanced Data Preprocessing for Intersites Web Usage Mining," *IEEE Intelligent Systems*, vol.19, Issue 2, pp.59-65, 2004.

[6] Jaideep Srivastava, Robert Cooley, Musund Deshpande, Pang-Ning Tan "Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data," *Exploration ACM SIGKDD*, 2000.

[7] R. Cooley, B.Mobasher, and J.Srivastava, "Web Mining : Information and Pattern Discovery on the World Wide Web," *9th IEEE International Conference on Tools with Artificial Intelligent*, 1997.

[8] Kibum Kim, John M. Carroll and Mary Beth Rosson, "An Empirical Study of Web Personalization Assistants : Supporting End-Users in Web Information Systems," *IEEE 2002 Symposia on Human Centric Computing Languages and Environments*, pp.60-62, 2002.

[9] Kun_Lung Lu, Charu C. Aggarwal and Philip S. Yu, "Personalization with Dynamic Profiler," *3rd International workshop on Advanced Issues of E-Commerce and Web-Based Information Sytems(WECWIS)*, pp.12-20, 2001.

[10] Juan Velásquez, Hiroshi Yasuda and Terumasa Aoki, "Combining the Web Content and Usage mining to Understand the Visitor behavior in a Web Site," *3rd IEEE International Conference on Data Mining*, pp.669-672. 2003.

[11] L. Catledge and J. Pitkow, "Characterizing Browsing Behaviors on the World Wide Web," *Computer Networks and ISDN Systems*, vol. 27, no. 6, 1995.

[12] George T. Wang, "Web Search With Personalization and Knowledge," *IEEE 4th International Symposium on Multimedia Software Engineering*, 2002.

[13] Kwangnam Kim, Heebyung Yoon, Hwa-Soo Kim "Design and Implementation of XML-based Indexing Algorithm Using Depth-First and Shortest Distance Between Nodes," *31th KISS spring conference*, vol. 31, no. 1, p.547-549, 2004.

저 자 소 개

김광남(KwangNam Kim)

1994년 : 육군사관학교(이학사)
2004년 : 국방대학교 전산정보학과 석사과정

관심분야 : 웹 마이닝, 데이터 마이닝, 멀티에이전트

Phone : 02-300-2138
E-mail : kma6471@hanmail.net

윤희병(Heebyung Yoon)

1983년 : 해군사관학교(이학사)
1986년 : 연세대학교 (공학사)
1991년 : 미국 해군대학원 전산공학(석사)
1998년 : 미국 Georgia Institute of Technology 전산공학(박사)
2002년~현재 : 국방대학교 전산정보학과 조교수

관심분야 : 웹 마이닝, 임베디드 소프트웨어, 에이전트 시스템, 모바일 웹 검색

Phone : 02-300-2138
E-mail : hbyoon@kndu.ac.kr

김화수(Hwa-Soo Kim)
1976년 : 해군사관학교(이학사)
1984년 : 미국 해군대학원 전산학(석사)
1990년 : 미국 Case Western Reserve University 전산학(박사)
1991년~2002년 : 국방대학교 전산정보학과 교수
2003년~현재 : 아주대학교 정보통신대학원 교수
2003년~현재 : 국방부장관 정책자문위원(정보화분야)
2004년~현재 : 감사원자문위원(행정 및 국가안보분야)

관심분야 : S/W 개발비용 산정/분석, 최신 정보기술관리, 인공지능 및 전문가시스템 개발

Phone : 02-319-4080p714
E-mail : ajhskim@ajou.ac.kr