클러스터 내의 문서 유사도를 이용한 추천 시스템
Recommendation System Using Document Similarity in Cluster

이재호, 서상훈(인터넷경영정보과)
Jae-Ho Lee, Sang-Hoon Suh (Dept. of Internet and Management Information System)

Key Words : Mining, ARHP, Browsing Pattern, Session, cluster, recommendation

ABSTRACT : Because web documents are created and purged rapidly, users require the recommend system that offers users to browse the web document with convenience and correctness. One largely untapped source of knowledge about large data collections is contained in the cumulative experiences of individuals finding useful information in the collection. Recommendation systems attempt to extract such useful information by capturing and mining one or more measures of the usefulness of the data. The existing Information Filtering system has the shortcoming that it must have user's profile. And Collaborative Filtering system has the shortcoming that users have to rate each web document first and in high-quantity, low-quality environments, users may cover only a tiny percentage of documents available. And dynamic recommendation system using the user browsing pattern also provides users with unrelated web documents. This paper classifies these web documents using the similarity between the web documents under the web document type and extracts the user browsing sequential pattern DB using the users' session information based on the web server log file. When user approaches the web document, the proposed Dynamic recommendation system recommends Top N-associated web documents set that has high similarity between current web document and other web documents and recommends set that has sequential specificity using the extracted informations and users' session information.

1. 서론

정제되지 않은 웹 데이터에는 사용자들의 축적된 경험들을 포함하는 유용한 정보들을 가지고 있다. 이러한 유용한 정보를 마이닝 기법이나 다른 측정 방법을 가지고 추출하여 추천 시스템을 통해 사용자에게 제공하려는 노력이 시도되고 있다[16].

기존의 정보 필터링(Information Filtering) 방식에 의한 추천 시스템은 항목 컨텐츠의 분석과 사용자들이 입력한 흥미로운 프로파일에 근거하여 사용자들의 특성을 파악하여 추천 서비스를 제공하려 하였다. 그러나 사용자들이 직접 입력한 정보는 매우 왜곡된 것일 수 있으며, 사용자 프로파일은 정적이기 때문에 시간이 지남에 따라 프로파일의 질적인 효과는 감소할 수밖에 없다는 문제점이 있다[11].

이러한 정보 필터링 방식에 의한 추천 시스템의 문제점을 개선하기 위해 사용자들로부터 먼저 웹 문서에 대한 평가를 입력받아 평가된 축적 정보를 다른 사용자들에게 제공하려는 협력적 필터링(Collaborative Filtering) 방식에 의한 추천 시스템이 제안되었다. 이러한 협력적 필터링 방식을 이용하는 추천 시스템에는 FireFly[3], GroupLens[8][14]와 같은 시스템들이 있다. 협력적 필터링 방식의 추천 시스템은 일정한 목적없이 웹 사이트를 방문한 사용자들에게는 미처 생각하지 못한 웹 문서들을 추천 문서로 제공받는다는 장점이 있다. 그러나 새로운 웹 문서가 출현할 때, 사용자들이 일정한 수 이상 평가되기 전까지는 다른 사용자들에게 추천 집합으로

제공되지 못하는 First Rater 문제, 대량의 웹 문서에서 사용자가 찾고자 하는 웹 문서가 상당히 드물 때 발생하는 Sparsity 문제 등의 다양한 문제점이 존재한다.

최근에는 이러한 단점을 보완하기 위해 협력적 필터링 방식에 컨텐츠를 적용하여 First Rater 문제나 Sparsity 문제를 해결하려는 연구가 있지만[14], 웹 문서들 사이의 연관성에 대한 고려가 여전히 미흡하다. 또한, 기존의 협력적 필터링 방식을 이용한 동적 추천 시스템이 사용자들의 입력 자료에 지나치게 의존하고 있는 문제점을 개선하기 위해 웹 마이닝 기법을 이용하여 사용자들의 브라우징 패턴 정보로부터 추천 문서를 제공하려는 연구도 시도되었다[5][13]. 그러나 이러한 추천 시스템은 사용자 브라우징 패턴을 분석하고 결정하기 위해 데이터 마이닝의 연관 규칙 알고리즘을 이용하고 있는데, 이러한 방법은 사용자들의 브라우징 순서를 고려하지 않고 단순히 빈번하게 동시에 발생하는 웹 문서들의 요청에 대해서만 규칙을 생성하기 때문에 시간상의 선후 관계가 존재하는 브라우징 패턴을 정확하게 분석하지 못하고, 웹 문서들의 내용 정보를 무시함으로써 웹 문서들간의 내용적 측면에서의 연관성을 고려하지 않는다.

또한 연관 웹 문서 분류와 브라우징 순차 패턴을 이용한 동적 링킹 시스템[18]에서는 WEBMINER 시스템[7]이 지닌 문제점을 해결하기 위해 웹 마이닝 기법 중 순차 패턴 알고리즘[2]을 이용하여 사용자들의 브라우징 순서에 대한 정보까지 고려하였다. 웹 문서들간의 연관성을 이용하기 위해서는 ARHP(Association Rule Hypergraph Partitioning) 알고리즘[12]을 이용하였지만, 추천 문서로 하이퍼링크 위주의 탐색 페이지와 같은 정보가 없는 단순 링크 기능을 지닌 불필요한 웹 문서까지 추천 문서로 제공하는 문제점이 있다.

본 논문에서는 웹 문서의 형식 결정 단계를 거친 후 탐색 페이지를 제외한 나머지 웹 문서들간의 유사도를 측정하여 사용자에게 웹 문서를 추천한다. 그리고 웹 서버 로그 파일에 포함된 유용한 정보들로부터 사용자들의 브라우징 순차 패턴을 생성, 웹 문서의 형식에 따라 연관된 웹 문서뿐만 아니라 사용자들이 자주 지나간 순차적인 특성을 가진 웹 문서를 추천 문서로 제공한다. 이때 추천 웹 문서 집합이 탐색 페이지이면 사용자 브라우징 순차 패턴 DB에서 사용자들이 자주 항해하는 순차적인 웹 문서 중 정보가 들어있는 내용 페이지를 사용자에게 최종 추천 문서로 제공한다.

2. 관련 연구 및 동향

2.1 추천시스템(Recommender System)

일반적으로 추천시스템은 전자상거래 사이트에서 고객에게 상품을 제안해 주거니 고객의 구매 과정에 도움이 될 수 있는 정보를 제공해 주는데 이용하고 있다. 이러한 추천 시스템은 고객이 원하는 상품까지 안내하기 위해 상품 지식을 이용하는데 이 지식은 전문가에 의해 정해지거나 고객의 행동으로부터 학습된 것을 이용한다.

추천시스템을 위한 연구는 협력적 여과(collaborative filtering)방법과 내용 기반 여과(content based filtering) 방법과 이들이 혼합된 방법들이 많이 이우러졌다. 내용기반 여과 방법의 문제점을 해결하고자 제시된 협력적 여과 방법에 대한 연구로써 GrpupLens[8]가 대표적인데, 추천을 위해 사용자들의 입력한 문서에 대한 평가를 이용해 비슷한 성향을 띠는 그룹내의 사람들의 행동을 관찰함으써 추천 서비스를 가능하게 하는 시스템이다. 일반적으로 협력적 여과 알고리즘은 목표 고객과 비슷한 이력을 가진 이웃(neighbor)으로 표현되는 고객 그룹을 찾는 통계 기법이다. 따라서 선호도가 비슷한 다른 고객들의 선호도를 바탕으로 고객에게 컨텐츠 및 서비스를 추천하는 것이다. 이는 유사한 선호도를 가진 고객들은 서로 각각 구매한 제품이 서로에

게 높은 구매율을 보이는 통계에 기초한 것이다. 즉 비슷한 취향을 가진 고객들에게 서로 아직 구매하지 않은 상품들을 교차 추천하거나 분류된 고객의 취향이나 생활형태에 따라 상품을 추천하는 형태로 서비스를 제공한다.

규칙 기반 필터링(rule-based filtering) 또한 추천 시스템에서 많이 쓰이는 방법인데, 웹 사이트에서 고객을 과거 거래 정보와 개인정보에 따라 분류하여 그에 따른 적절한 서비스 및 컨텐츠를 추천하는 것이다. 이 방법은 웹 사이트 운영 방침에 따른 컨텐츠를 제공한다는 점에서 규칙 기반 매칭(rule-based matching)이라고 부르며, 사용자에게 신상정보, 관심정보, 선호도 등에 대한 몇 가지 질문을 하는 것이 일반적이다. 몇가지 질문을 거침으로서 그 사람에 대한 정보들의 프로파일이 생기게 되는데, 웹 개인화 담당자는 이렇게 수집된 사용자의 인구통계학적, 심리적 정보와 사용자의 선호도 정보에 알맞게 정보 및 상품을 추천 또는 제공하는 것이다.

지능형 에이전트(intelligent agent)는 자율성, 추론 커뮤니케이션을 수행하며, 네트워크 상에 존재하는 개념으로 '자동학습을 통한 추론"으로 나타낼 수 있다. 학습 에이전트 기술은 사용자가 사이트 내에서 어떤 페이지에 오래 머무르는지, 어떤 페이지를 인쇄하는 지, 어떤 제품을 구매하는 지 등과 같은 사용자들의 행동을 기준으로 사용자의 선호도와 관심을 알아내고 이를 바탕으로 사용자들에게 적절한 내용을 제공한다.

최근에는 데이터 마이닝(data mining) 기법을 이용해서 웹상에서의 사용자들의 이용 패턴(usage pattern)을 찾아냄으로써 개인화된 웹 사이트 구축 및 추천 시스템에 대한 연구가 활발히 진행되고 있다.

이러한 데이터 마이닝 기법을 이용하여 적절한 추천 서비스를 하기 위해서는 사용자에 대한 정보가 절대적인 가치를 가지는데, 사용자들의 웹 서핑 동안의 정보를 트래킹(tracking)하여 얻는 방법으로 대표적인 것이 웹 로그 분석이다.

3. 로그분석과 문서 유사도를 이용한 동적 추천 시스템

WEBMINER 시스템[7]은 웹 서버에 기록된 로그 파일을 이용하여 연관 규칙과 웹 사이트상의 문서들간의 거리를 가지고 추천 문서를 생성한다. 그리고 연관 웹 문서 분류와 순차 패턴을 이용한 추천 시스템[18]은 추천을 위해 순차 패턴과 연관 웹 문서를 이용한다. 그러나 위의 두 시스템들은 의미가 없는 정보만을 가지고 있는 웹 문서로의 연결 역할만 하는 웹 문서까지 추천 집합으로 제공하고 있다.

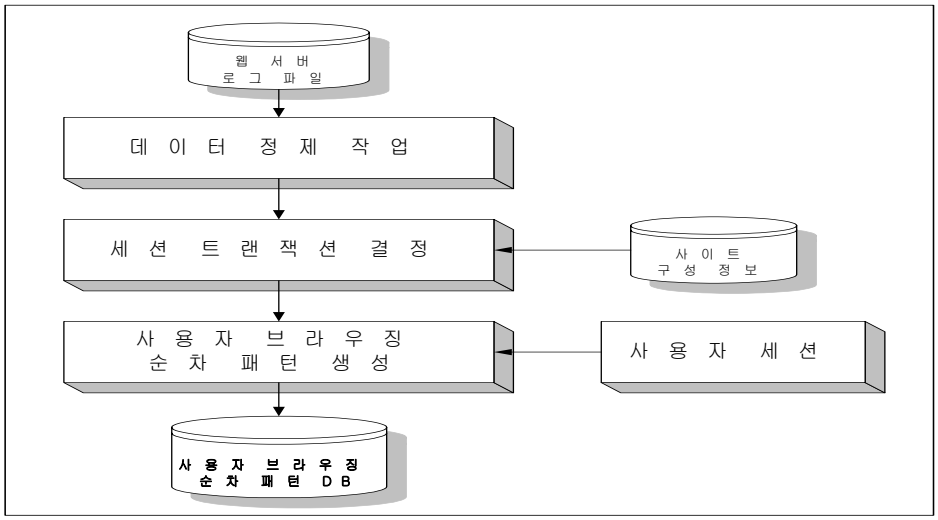
본 논문에서 설계한 시스템은 첫째, 사용자 브라우징 순차 패턴 DB를 구축하기 위해 웹 서버에 저장되어 있는 로그 파일로부터 데이터 정제 작업을 수행한 후, 세션 트랜잭션을 생성하고 사용자 브라우징 순차 패턴을 생성한다. 둘째, 웹 사이트에 있는 웹 문서들을 대상으로 웹 문서 형식을 결정하고 연관 웹 문서 분류를 수행하기 위해 ARHP 알고리즘을 이용한 웹 문서들을 군집화하고 자카드 계수를 이용하여 웹 문서들을 분류한다.

두 과정에서 추출된 사용자 브라우징 순차 패턴 DB와 연관 웹 문서 정보를 이용하여 사용자가 웹 문서에 접근했을 때 동적 추천 알고리즘을 이용하여 사용자에게 연관된 웹 문서나 먼저 방문한 다른 사용자들의 순차적 경험이 내포된 웹 문서를 추천 문서로 제공한다.

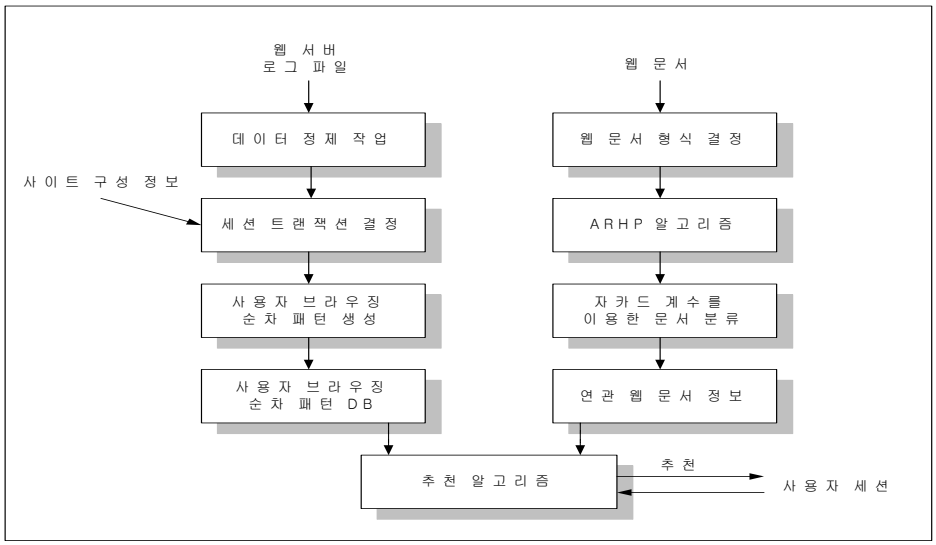
Fig.1은 본 논문에서 설계한 동적 추천 시스템에 대한 전체 구성도이다.

3.1 사용자 브라우징 순차 패턴 분석

사용자 브라우징 순차 패턴 정보를 생성하기 위해 웹 서버의 로그 파일로부터 사용자들의 웹 서버 접근 기록을 얻어낸다. 웹 서버의 로그 파일에는 분석에 필요하지 않은 다량의 정보도 포함되어 있기 때문에 먼저 로그 파일에 대한 정제 작업을 수행한다. 또한 세션 트랜잭션 정보를 얻기 위해 사이트 구성 정보를 참조한다. 그리고 세션 트랜잭션에 대해 순차 패턴 알고리즘을 이용하여 사용자 브라우징 순차 패턴을 생성하고, 이 정보로부터 사용자 브라우징 순차 패턴 DB를 생성한다. 사용자 브라우징 순차 패턴 분석은 Fig.2와 같이 진행된다.



(Fig.1) A concept diagram of user browsing sequence patter



(Fig.2) A composition diagram of dynamic recommendation system

3.1.1 데이터 정제 작업

사용자 브라우징 순차 패턴 정보를 추출하기 위해 가공되지 않은 원시 데이터인 웹 서버의 로그 파일에서 불필요한 정보들을 제거하는 데이터 정제 작업을 수행해야 한다. 사용자가 웹 서버에 접근하여 웹 문서를 요청하면 웹 로그 파일에는 그 웹 문서 안의 여러 파일들의 요청

까지 자동으로 로그 파일에 기록된다. 따라서 분석하려는 자료의 성향에 따라 하나의 웹 문서를 대표할 수 있는 웹 문서 파일을 제외한 다른 파일들은 로그 파일로부터 제거한다.

본 논문에서는 인하대학교 대학원 웹 사이트에 대한 동적 추천 시스템을 구축하기 위한 것으로, 사용자가 요청한 파일의 확장자가 "*.htm", "*.html"인 파일을 제외한 모든 기록들을 제거하고, 사용자의 IP 주소, 요청 시각, 요청 URL 필드만을 남긴 나머지 데이터는 제거한다. 그리고 정제된 로그 파일로부터 세션 트랜잭션 결정을 위한 항목 $I=\{IP, TIME, URL\}$을 생성한다. 여기서 IP는 사용자가 웹 서버에 접근했을 때 사용된 컴퓨터의 IP 주소이고, $TIME$은 사용자가 웹 문서를 요청한 시간, URL은 사용자가 웹 서버에 요청한 웹 문서의 URL이다.

3.1.2 사이트 구성 정보 복원 작업

사용자들이 사용하는 브라우저의 로컬 캐쉬와 프록시 서버의 캐쉬 사용으로 인해 요청되지 않은 기록을 복원하기 위한 사이트 구성 정보가 필요한데, 이를 Path Completion이라 한다[10]. 본 논문에서는 제약 사항으로 각 사용자는 프록시 서버를 사용하지 않고, 고정 할당 IP 주소를 사용하고 있다고 가정한다.

3.1.3 세션 트랜잭션 결정

사용자 브라우징 순차 패턴을 추출하기 위해서는 요청된 웹 문서들을 항목으로 하는 세션 트랜잭션이 결정되어야 한다. 본 논문에서는 한 사용자가 한 번의 방문동안 요청한 웹 문서들에 대응하는 항목 I들의 집합으로 세션 트랜잭션을 구성한다. 세션 트랜잭션을 결정하기 위한 방법에는 Reference Length Module, Maximal Forward Reference Module, Time Window Module 등이 있다[6][7]. 본 논문에서는 대부분의 상용 제품에서 사용되는 Time Window Module을 이용하고 시간의 한계값은 30분을 사용한다[7]. 즉 사용자가 웹 서버에 30분 동안 아무런 요청 기록이 없으면 세션의 종료로 가정한다. 세션 트랜잭션 ST는 다음과 같이 정의된다[18].

$$ST_k=[\{IP,\ TIME_1,\ URL_1\},\ldots,\{IP,\ TIME_i,\ URL_i\},\ldots,\{IP,\ TIME_n,\ URL_n\}]$$

여기서, 임의의 k 세션 트랜잭션(ST_k)에 포함된 IP는 모두 동일하고, $TIME_{i+1}-TIME_i$는 30분 미만이어야 한다. 만일 30분 이상이 되면 세션 트랜잭션은 종료되었다고 간주하고 새로운 세션 트랜잭션을 생성한다.

3.1.4 사용자 브라우징 순차 패턴 생성

세션 트랜잭션이 결정되면 세션 트랜잭션들 중에서 URL들만을 항목으로 하는 URL 트랜잭션(UT)을 다음과 같이 구성한다.

$$UT_k = \{\ URL_1, URL_2, \ldots, URL_n\}$$

URL 트랜잭션을 대상으로 AprioriAll 알고리즘[2]을 이용하여 사용자 브라우징 순차 패턴을

생성하며, 생성된 Large 항목집합을 Large 시퀀스라고 부른다. 이때 Large 시퀀스 자체가 얻고자 하는 순차 패턴이다. 위의 방법을 통해 얻어진 Large 시퀀스는 사전에 정의된 최소 지지도를 만족하는 순차 패턴을 의미하는데, 이 순차 패턴에 포함된 *URL*들은 "최소 지지도를 만족하면서 사용자들이 한번의 방문(세션)동안 순차적으로 방문한 웹 문서들"이라는 의미를 지닌다.

3.2 웹 문서 형식 결정

웹 문서는 형식에 따라 주요 페이지(Head page), 내용 페이지(Content page), 탐색 페이지(Navigation page), 참조 페이지(Look-up page), 개인 페이지(Personal page) 등으로 분류할 수 있다[4]. 주요 페이지는 사용자가 웹 사이트를 방문하였을 때의 첫 번째 페이지이고, 내용 페이지는 웹 사이트가 제공하는 정보의 내용이 포함되어 있는 페이지이고, 탐색 페이지는 웹 문서 내의 하이퍼링크를 통해 내용 페이지로 안내하는 페이지이다. 그리고 참조 페이지는 정의와 약어 표현을 위한 페이지이고, 개인 페이지는 각 개인들의 특성을 지닌 정보가 들어있는 페이지이다. 본 논문에서는 웹 문서의 형식 중 주요 페이지와 참조 페이지는 특성상 탐색 페이지에 포함시켰고, 개인 페이지는 실험 대상에 존재하지 않기 때문에, 웹 문서 형식을 탐색 페이지와 내용 페이지만으로 분류한다.

웹 문서의 형식 결정 기준은 문서 내의 단어 수와 링크 수, 단어들간의 유사도를 측정하여 판별한다. 웹 문서에 출현하는 단어를 이용하여 웹 문서에 출현하는 단어의 수가 한계값 이상이면 이를 내용 페이지로 분류한다. 이때 한계값을 낮추면 재현율을 높일 수 있다. 그리고 다른 웹 문서와 연결된 하이퍼링크가 한계값 이상이면 정보가 없는 탐색 페이지로 추정한다. 또한 단어들간의 유사도를 이용하여 하나의 웹 문서에 나타나는 단어들 사이의 유사도가 높으면 내용이 있는 내용 페이지로 추정한다. 즉, 단어들 사이의 유사도가 높으면 그 웹 문서는 내용 페이지이고 유사도가 낮으면 내용이 없는 탐색 페이지로 간주하며, 웹 문서 형식 결정 기준은 Table 1을 따른다.

Table 1. Decision Criterion of Web Document Format

	단어 수 (*Doc_N*)	링크 수 (*Link*)	단어들 간의 유사도 (*Word_Sim*)
내용 페이지	$Doc\_N \geq \alpha$	$Link < \beta$	$Word\_Sim \geq \gamma$
탐색 페이지	$Doc\_N < \alpha$	$Link \geq \beta$	$Word\_Sim < \gamma$

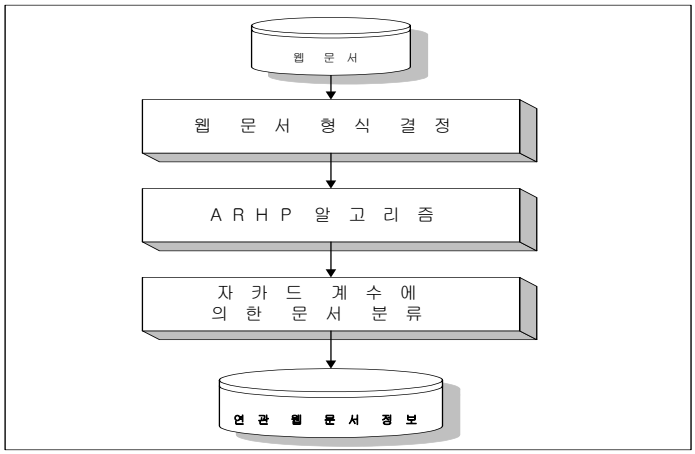
Table 1에서 단어 들간의 유사도를 측정하기 위해서는 단어들의 연관성을 정량적으로 나타내는 상호 정보량[17]을 이용한다.

3.3 연관 웹 문서 분류

연관 웹 문서 분류를 위한 단계적인 흐름은 Fig.3과 같이 진행된다.

연관 웹 문서 분류를 수행하기 위해 웹 문서 형식이 내용 페이지인 웹 문서들을 대상으로 각 문서에서 출현하는 단어들을 이용하여 동일한 식별자를 갖는 하나의 트랜잭션을 구성한다.

그리고 연관 규칙 알고리즘을 이용하여 단어들간의 연관 규칙을 생성하고, 생성된 연관 규칙의 신뢰도를 가중치로 사용하여 ARHP 알고리즘을 적용, 여러 웹 문서에서 동시에 출현하는 명사들의 리스트를 군집화한다[18].



(Fig.3) A concept diagram of a related web document

그리고 각 클러스터 내에 포함된 단어 리스트를 가지고 벡터 모델, $V_{Cluster_i} = (w_{i1}, w_{i2}, \ldots, w_{ij}, \ldots, w_{in})$로 표현한다. 본 논문에서는 웹 문서의 특징을 추출하기 위해 역문헌 빈도를 이용하는 것이 아니라, 처리 속도와 정확도를 높이기 위해 웹 문서에서 추출된 모든 명사들을 이용한다. 따라서 확률 모델보다는 벡터 모델이 적합하다.

벡터 모델에서 w_{ij}의 값은 이진수로 표현되며, 단어가 $Cluster_i$에 포함되면 1로 표현하고 아니면 0으로 표현한다. 분류될 웹 문서를 위와 같이 벡터 모델로 변환시킨 후 유사도를 측정한다. 유사도 측정은 (1)의 자카드 계수[16]를 이용한다. 자카드 계수를 이용하는 이유는 유사도 측정 공식 중 자카드 계수가 문서 클러스터링에 사용되는 가장 보편적인 함수이기 때문이다.

$$Sim(V_i , V_j) = \frac{\sum_{k=1}^{n} (w_{ik} \cdot w_{jk})}{\sum_{k=1}^{n} w_{ik} + \sum_{k=1}^{n} w_{jk} - \sum_{k=1}^{n} w_{ik} \cdot w_{jk}} \qquad (1)$$

여기서 n은 $Cluster_i$에 포함된 단어의 수이고, w_{ik}는 벡터 V_i의 k번째 단어의 값이다[9].

(1)의 $Sim(V_i, V_j)$ 공식은 두 번 사용되는데, 첫째는 클러스터 내에 포함된 단어 리스트들을 가지고 전체 웹 문서들과의 유사도를 측정하여 가장 큰 값을 갖는 문서를 각 클러스터에 분류할 경우에, 그리고 두 번째로는 사용자들이 웹 문서에 접근했을 때 각 클러스터 내에 포함되어 있는 웹 문서들과 현재 접근중인 웹 문서 사이의 유사도를 측정하여 추천할 집합을 찾을 경우에 사용한다. 이때, 첫 번째 경우는 $V_i = (1, 1, \ldots, 1)$과 웹 문서 V_j의 벡터를 이용하며, $Sim(V_i, V_j)$ 값들 중 가장 큰 클러스터에 웹 문서를 할당하여 연관 웹 문서를 분류한다. 그리고 두 번째 경우는 현재 접근중인 웹 문서 V_i와 클러스터내에 포함된 n개의 웹 문서들 사이의 $Sim(V_i, V_j)$를 측정하여 상위 N개의 추천 집합을 생성한다.

3.4 동적 추천 알고리즘

동적 추천 알고리즘은 사용자 브라우징 순차 패턴으로부터 현재 사용자의 세션을 갖는 추천 집합과 사용자가 방문하고 있는 현재 웹 문서와 가장 연관된 웹 문서를 제공할 추천 집합을 생성하는 알고리즘으로 본 논문에서 설계한 동적 추천 알고리즘은 Algorithm.1과 같다.

Algorithm.1에서 사용자 세션 중 last_url의 웹 문서 형식이 탐색 페이지인지 내용 페이지인가에 따라 추천 집합을 생성한다. 사용자의 last_url의 웹 문서 형식이 내용 페이지이면, 현재 세션을 포함하면서 1이 더 큰 large 시퀀스 집합 중에서 추천 집합을 생성한다. 만약 last_url의 웹 문서 형식이 탐색 페이지이면, 사용자 브라우징 순차 패턴 DB에서 last_url을 포함하면서 최소 지지도 이상을 가진 large 시퀀스 집합에서 추천 집합을 생성한다. 이때 동적 추천 알고리즘에서 사용하는 신뢰도($session \Rightarrow url$)는 다음 (2)와 같다.

$$\text{신뢰도}(session \Rightarrow url) = \frac{| session \cap url |}{| session |} \quad | \qquad (2)$$

신뢰도($session \Rightarrow url$)는 추천 집합의 순위 결정을 위한 가중치로 사용되며, 최소 신뢰도를 만족하는 url들만을 추천 집합에 포함시킨다. 이때 추천될 url의 웹 문서 형식이 탐색 페이지이면 Seq 함수로부터 반환된 웹 문서를 포함시킨다. Seq 함수는 url을 포함하는 순차 패턴 DB에서 사용자의 세션을 가지고 사용자들이 자주 항해하는 순차적인 특성을 가진 웹 문서를 반환하는 함수이다. 예를 들어, A, C문서는 내용 페이지이고, B문서는 탐색 페이지라고 가정할 때, 만약 A→B→C로 항해하는 사용자가 많다고 가정하면, 동적 추천 시스템은 사용자가 A문서를 방문했을 때 B문서를 추천하는 것이 아니라 C문서를 추천 집합에 포함시켜 제공하는 것이다. 이것은 Algorithm.1에 의해 B문서가 웹 문서 형식이 탐색 페이지이기 때문에 C문서를 제공하는 것이다. 또한 last_url을 가진 클러스터내의 웹 문서들 사이의 유사도가 높은 상위 N개의 웹 문서들을 추천 집합에 포함시킨다.

4. 실험 및 결과

본 논문의 실험 환경으로는 Windows NT 서버 4.0을 사용하였으며, 시스템의 구현을 위해 MS Visual C++ 6.0을 사용하였다. 실험 데이터로는 인하대학교 대학원 홈페이지를 서비스하는 웹 서버의 Common Logfile Format(CLF) 로그 파일 가운데, 2002년 5월 18일부터 10월 2일까지의 기록된 정보를 이용하였다. 이때 사용된 데이터는 정제된 로그 파일이다. 그리고 인하대학교 대학원 웹 문서 188개 중 연관 웹 문서 분류를 위해 웹 문서 형식이 내용 페이지인 169개의 웹 문서들을 이용하였으며, 정보가 없는 하이퍼링크 위주의 웹 문서들은 연관 웹 문서 분류에서 제거된다.

사용자 브라우징 순차 패턴을 생성하기 위한 측정 기준으로 지지도의 최소 지지도 한계값은 30%로 설정하였다. 웹 서버 로그 파일로부터 순차 패턴 알고리즘[(2)]을 이용하여 최소 지지도 30%를 만족하는 사용자 브라우징 순차 패턴을 생성하였으며, 이때 추출된 순차 패턴의 수는 481개이다.

Table 2는 추출된 사용자 브라우징 순차 패턴에 대해 지지도를 기준으로 정렬한 데이터베이스의 일부분을 나타낸다. Table 2의 왼쪽 필드는 사용자 브라우징 순차 패턴들이고, 오른쪽 필드는 각 순차 패턴에 대한 지지도를 나타낸다.

(Algorithm.1) Dynamic recommendation Algorithm

```
Input : cur_session // 현재 사용자 세션
        last_url    // 사용자가 가장 최근에 요청한 URL
        σ           // 최소 지지도
        α           // 최소 신뢰도
        ξ           // 유사도 한계값

Output : Recommend1 // 순차 패턴에 포함된 추천 문서 집합
         Recommend2 // 연관 문서에 포함된 추천 문서 집합

Recommend1 = Recommend2 = ∅ ;
if ( last_url.type == content_page ){
for each I do   // I는 cur_session을 포함하는 size가 |cur_session|+1
                // 인 large sequence 집합
  if ( 지지도(I) >= σ )
     confidence = 신뢰도(session⇒url) ;
     if ( confidence >= α ){
       url.score = confidence ;
       if ( url.type == navigation_page )
         url = Seq(url) ;
       Recommend1 += url ;
      }
  for each url do
     if (( value = Sim(last_url, url)) >= ξ )
       Recommend2 += url;
 }
else if ( last_url.type == navigation_page ){
    while (( url = Seq(last_url)) >= σ )
       if ( url.confidence >= α ){
       url.score = url.confidence ;
       if ( url.type == navigation_page )
         url = Seq(url) ;
       Recommend1 += url ;
      }
 }
```

Table 2. sequential pattern of user browsing

사용자 브라우징 순차 패턴	지지도 (%)
[/~grad/]→[/~grad/menu.htm]	51.78
[/~grad/]→[/~grad/sugang/2002-1iljung.hwp]	50.23
[/~grad/]→[/~grad/sugang/main.htm]	47.46
[/~grad/]→[/~grad/sugang/main.htm]→[/~grad/menu.htm]	27.52
[/~grad/]→[/~grad/notice/2002-1enter/2002-1_grad.html]	24.85
[/~grad/]→[/~grad/labs/main.htm]→[/~grad/labs/eng.htm]	23.85
[/~grad/sugang/index.html]→[/~grad/sugang/main.htm]	20.34
...	...

웹 문서 형식 결정을 위한 한계값으로 문서내의 단어 수와 링크 수는 각각 150개와 25개로 설정하였고, 단어들 간의 연관도는 상호 정보량의 평균값을 이용하였으며, 이때 한계값은 0.15로 설정하였다.

웹 문서 형식이 내용 페이지인 169개의 웹 문서를 대상으로 문서에 빈번히 출현하는 단어들

의 집합으로부터 최소 지지도 30%, 최소 신뢰도 50%를 만족하는 단어들의 쌍을 연관 규칙 알고리즘[1]을 이용하여 추출한다. 이때 ARHP 알고리즘에 적용하기 위해 연관 단어들의 신뢰도를 가중치로 하여 5개의 단어 클러스터를 생성하였다. 각 단어 클러스터에 포함된 단어들을 이용하여 169개의 웹 문서를 분류하며, 이때 각 클러스터에 포함된 단어들을 벡터로 표기하고, 각 문서들과의 유사도를 측정하기 위해 자카드 계수 (1)를 이용하여, 유사도의 크기가 가장 큰 클러스터에 웹 문서를 할당한다. 그 결과 클러스터 번호 1, 2, 3, 4, 5에는 각각 73, 26, 13, 45, 12개의 웹 문서들이 할당되었다.

Table 3은 분류된 연관 웹 문서의 벡터 표현과 해당 클러스터 내에 포함된 웹 문서 수를 나타낸다. 분류된 연관 웹 문서 정보는 사용자가 웹 서버에 접근했을 때, 현재 웹 문서와 클러스터 내의 문서들과의 유사도를 측정하여 방문하고 있는 웹 문서와 가장 유사한 문서를 찾아내어 추천 집합으로 사용자에게 제공한다. 이렇게 생성된 순차 패턴 DB와 연관 웹 문서 정보는 사용자가 웹 페이지에 접근했을 때의 입력 자료로 사용된다.

Table 3. vector expression of web document classified and associated

번호	클러스터된 단어 및 연관 문서	문서수
1	[개설][입학][자격][전공][전형][제출][과목][교수] [구술][학과][학문][합격][특별][학부][학술지]...	73
	[~grad/sugang/2002-1orientation.html] (1,1,1,1,0,1,1,0,0,0,1,1,1,1,0,1,0,1,0, ... ,0,1,1) [/~grad/sugang/2002-1iljung.hwp] (1,1,0,1,1,1,0,0,1,0,1,0,1,1,1,1,0,1,0, ... ,0,1,0)...	
2	[정보][정의][정책][제도][계산][서류][설명][교육] [공동][공통][과학][동일][문학][발표][지원]...	26
	[~grad/about/cooperation.htm] (0,0,1,1,1,1,1,1,1,1,1,1,1,0,0,0,1,1,1, ... ,0,1,1) [~grad/notice/2002-1jaguekhab.html] (1,1,1,0,1,1,0,1,0,1,0,1,1,0,0,0,1,1,0, ... ,1,1,1)...	
3	[경제][관리][기관][흐름][모델][목표][문화][물리] [반응][분류][사례][발달][운영][파악][정론]...	13
	[~grad/labs/sci.htm] (0,1,0,0,0,0,1,0,0,1,1,0,1,0,0,1,0,0,0, ... ,0,0,1) [~grad/course/degree/sangsan.htm] (0,1,1,0,1,0,0,1,1,1,0,1,0,0,1,0,0,0,1, ... ,1,1,0)...	
4	[강의][개념][개발][개요][검토][제반][고찰][과정] [소개][동향][교과][관계][습득][원리][주제]...	45
	[~grad/course/degree/bio.htm] (1,1,1,1,0,1,1,1,0,1,1,1,1,0,1,1,0,1,1, ... ,1,1,1) [~grad/course/degree/교육-1.htm] (1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1, ... ,1,1,1)...	
5	[공정][공학][관점][세미나][기술][시스템][목적][제어][발전][효율][컴퓨터][자료][수행][설계][에너지]...	12
	[~grad/labs/eng.htm] (1,1,1,0,1,0,1,0,1,0,1,0,1,0,0,1,1,1,1, ... ,1,0,0) [~grad/about/student.htm] (0,0,1,0,0,1,0,1,0,1,0,0,0,0,0,1,1,0,0, ... ,0,1,0)...	

Table 4는 수강 신청 웹 문서를 방문했을 때, 세 개의 추천 시스템이 추천한 상위 12개의 웹 문서를 보여준다. 세 개의 추천 시스템은 WEBMINER, 연관 웹 문서 분류와 브라우징 순차 패

턴을 이용한 동적 링킹 시스템(Dynamic Linking System Using Related Web Documents Classification and Browsing Sequential Patterns; DLS), 그리고 본 논문에서 제안한 동적 추천 시스템(Dynamic Recommendation System Using User Sequential Patterns and Document Similarity in Cluster; DRS)이다.

Table 5는 Table 4를 이용하여 상위 문서 4, 8, 12개의 문서를 대상으로 정확도, 재현율, F-measure[15]를 계산한 값을 나타낸다. 여기서, F-measure는 정확도나 재현율을 결합하여 같은 가중치를 부여함으로써, 질적인 판단을 높이기 위해 이용하는데, 측정 방법은 (3)과 같다.

$$\text{정확도} = \frac{\text{검색된 적합 문서 수}}{\text{상위 } N \text{ 문서 수}}$$

$$\text{재현율} = \frac{\text{검색된 적합 문서 수}}{\text{테스트 문서 총 수}} \qquad (3)$$

$$F-measure = \frac{2 \times \text{정확도} \times \text{재현율}}{(\text{정확도} + \text{재현율})}$$

Table 4. the comparison of Top-12 recommended web document

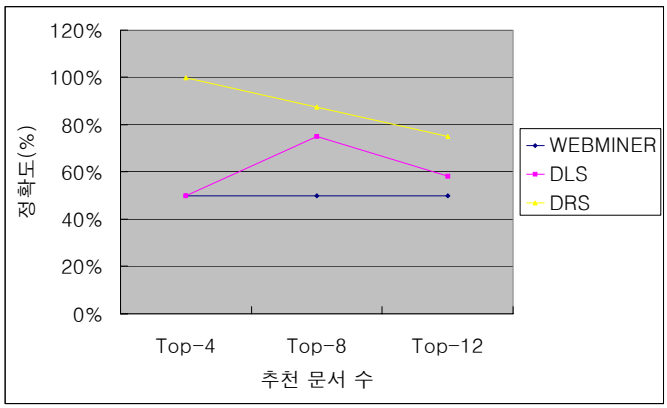
	WEBMINER	DLS	DRS
Top-12	[/~grad/]	[/~grad/]	[/~grad/sugang/2002-1iljung.hwp]
	[/~grad/menu.htm]	[/~grad/sugang/2002-1iljung.hwp]	[/~grad/sugang/2002-1orientation.html]
	[/~grad/sugang/2001-1iljung.hwp]	[/~grad/menu.htm]	[/~grad/sugang/2002-1gangjwa.html]
	[/~grad/sugang/2001-1orientation.html]	[/~grad/sugang/2002-1orientation.html]	[/~grad/notice/2002-1enter/2000-2_deonglok.htm]
	[/~grad/sugang/2001-1gangjwa.html]	[/~grad/sugang/2002-1gangjwa.html]	[/~grad/course/degree.htm]
	[/~grad/course/degree.htm]	[/~grad/notice/2002-1enter/2000-2_deonglok.htm]	[class.inha.ac.kr/lecture/grd_lectime/index.htm]
	[/~grad/notice/main.htm]	[/~grad/course/degree.htm]	[/~grad/notice/2002-1enter/2002-1_grad.html]
	[/~grad/labs/eng.htm]	[class.inha.ac.kr/lecture/grd_lectime/index.htm]	[/~grad/notice/2002-1enter/2002-1_4grad.html]
	[/~grad/notice/2001-1enter/2000-2_deonglok.htm]	[/~grad/notice/2002-1enter/2001-1_grad.html]	[/~grad/notice/9.10ganghakkeum.htm]
	[class.inha.ac.kr/lecture/grd_lectime/index.htm]	[/~grad/notice/9.10ganghakkeum.htm]	[/~grad/course/degree/전자계산공.htm]
	[/~grad/course/degree/전자계산공.htm]	[/~grad/course/degree/전자계산공.htm]	[/~grad/course/degree/전자공.htm]
	[/~grad/course/degree/전자공.htm]	[/~grad/course/degree/전자공.htm]	[/~grad/susik/main2.htm]

Table 5. the comparison of Top-N efficiency

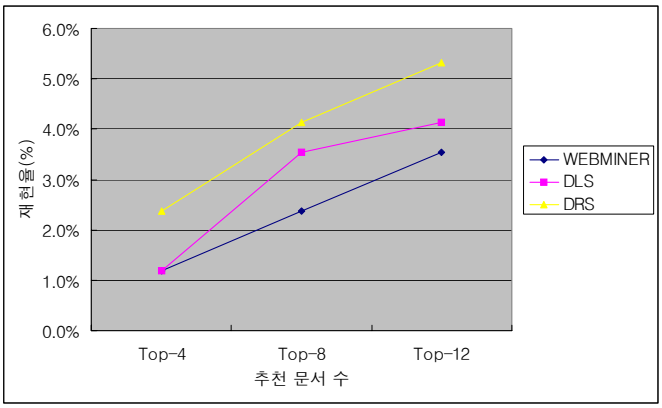
	정확도(%)			재현율(%)			F-measure		
	WEBMINER	DLS	DRS	WEBMINER	DLS	DRS	WEBMINER	DLS	DRS
Top-4	50.0	50.0	100.0	1.2	1.2	2.4	0.023	0.023	0.046
Top-8	50.0	75.0	87.5	2.4	3.6	4.1	0.045	0.068	0.079
Top-12	50.0	58.3	75.0	3.6	4.1	5.3	0.066	0.077	0.099

Fig.4와 Fig.5는 추천될 상위 문서의 수(N)를 4, 8, 12로 설정하였을 때의 정확도와 재현율 관계를 나타낸다. WEBMINER 시스템은 연관 규칙을 이용하여 추천 문서를 제공하기 때문에, [/~grad], [/~grad/menu.htm], [/~grad/labs/eng.htm] 등과 같이 빈번히 출현하지만 아무런 정보가 없는 웹 문서들을 제공한다. DLS 시스템은 순차 패턴과 연관 웹 문서를 이용하여 추천 문서를 제공하기 때문에 [/~grad/notice/2002-1enter/2002-1_grad.html] [/~grad/notice/9.10ganghakkeum.htm] 등과 같은 연관 문서가 포함되지만 WEBMINER 시스템처럼 정보가 없는 탐색 페이지를 추천 문서로 제공한다. 본 논문에서 제안한 DRS 시스템은 추천 문서를 웹 형식에 따라 다르게 제공하기 때문에, WEBMINER, DLS 시스템에서 제공되는 정보가 없는 문서, 즉 탐색 페이지를 추천 대상에서 제외시킴으로써, Top-4, Top-8, Top-12에서 정확도가 WEBMINER보다 50%, 37.5%, 25%, 그리고 DLS 보다는 각각 50%, 12.5%, 16.7% 향상되었다. 또한 재현율에서도 WEBMINER보다 1.2%, 1.7%, 1.7%, 그리고 DLS 보다는 1.2%, 0.5%, 1.2% 향상되었다.

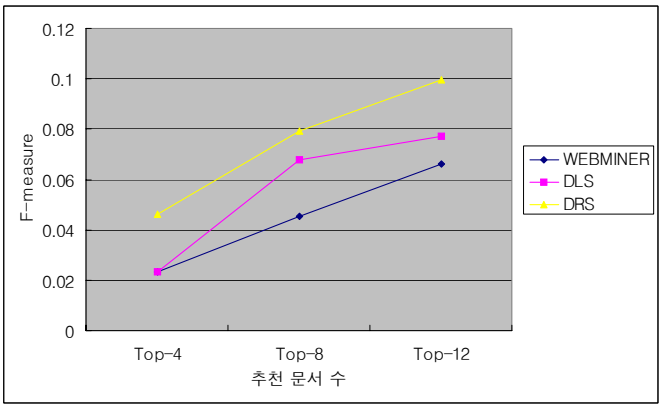
Fig.6은 각각 상위 N개 문서의 F-measure를 나타내는 것으로, 실험결과 본 논문에서 제안하는 DRS 시스템이 F-measure에서도 Top-4, Top-8, Top-12에서 WEBMINER 보다 0.023, 0.034, 0.033, 그리고 DLS 보다 0.023, 0.011, 0.022 향상되었다. 그 이유는 제안하는 시스템이 추천하는 문서는 사전 정보를 갖고 있지 않는 탐색 결과 페이지들에 대해서는 추천 문서로 제공하지 않기 때문이다.



(Fig.4) The correctness rate of top N documents



(Fig5) the reappearance rate of top N documents



(Fig.6) F-measure of of top N documents

5. 결론

본 논문에서는 기존 추천 시스템의 문제점을 개선하기 위해, 사용자 브라우징 패턴 분석 시 데이터 마이닝 알고리즘의 순차 패턴을 사용하여, 사용자들의 웹 사이트 방문에 대한 순차적 특성을 고려하였다. 이때 사용자의 현재 세션과 가장 유사한 사용자 브라우징 순차 패턴 정보를 추천 문서로 제공하였다. 또한 탐색 페이지와 같이 정보가 없고 단순히 내용 페이지로의 하이퍼링크만을 제공하는 웹 문서를 추천 문서에서 제외시킴으로써 사용자들이 보다 편리하고, 정확하게 사이트를 브라우징할 수 있는 시스템을 설계하였다. 이때, 사용자들에게는 이전 사용자들의 사전 경험과 연관된 웹 문서들을 이용하여 연관문서를 추천하기 때문에 시스템의 성능을 높이고 사용자가 원하는 웹 문서를 빠르게 제공할 수 있다.

실험 결과 Table 6과 같이, 사용자의 연관 규칙이나 순차 패턴만을 사용하는 것보다 연관 웹 문서를 분류하고 추천된 문서들 중 정보가 없는 웹 문서를 필터링함으로써 향상된 정확도와 재현율, 그리고 F-measure 값을 얻을 수 있었다.

향후, 각 사용자의 개별화를 통해 문서를 군집화하고 사용자가 웹 문서를 방문하였을 때 비슷한 관심을 갖고 있는 사용자들로부터의 정보를 얻어 제공하는 연구가 필요할 것으로 판단된다. 또한 제공된 추천 집합이 사용자들에게 얼마나 적용되었는지 로그 파일의 재분석을 통해 추천 알고리즘이나 전처리 부분을 재조정할 필요가 있다. 그리고 사용자 브라우징 순차 패턴 DB 상에 포함되지 않은 예상치 못한 상황이 발생하였을 경우를 대비하여 사용자 브라우징 순차 패턴 집합의 기계 학습을 통해 추천 집합을 사용자에게 제공해야 할 필요가 있을 것으로

판단된다

Table 6. the comparison of improvement of Top-N efficiency

	향상된 정확도(%)		향상된 재현율(%)		향상된 F-measure	
	WEBMINER	DLS	WEBMINER	DLS	WEBMINER	DLS
Top-4	50	50	1.2	1.2	0.023	0.023
Top-8	37.5	12.5	1.7	0.5	0.034	0.011
Top-12	25	16.7	1.7	1.2	0.033	0.022

참고 문헌

(1) R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules," Proc. of the 20th VLDB Conference, pp. 487-499, 1994.

(2) R. Agrawal and R. Srikant, "Mining Sequential Patterns," Proc. of the Int'l Conference on Data Engineering (ICDE), Taipei, Taiwan, March 1995.

(3) U. Shardanand and P. Maes, "Social information filtering : algorithms for automating 'word mouth'," Proc. of ACM CHI Conference, 1995.

(4) R. Agrawal, et al., "The Quest Data Mining System," Proc. of the 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, August 1996.

(5) Tak Woon Yan, et al., "From user access patterns to dynamic hypertext linking," Computer Networks an ISDN Systems 28, pp.1007-1014, 1996.

(6) J. S. Park, et al., "Using a Hash-Based Method with Transaction Trimming for Mining Association Rules," TKDE 9(5), pp. 813-825, 1997.

(7) R. Cooley, et al., "Web Mining : Information and Pattern Discovery on the World Wide Web," Proc. of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997.

(8) J. Konstan, et al., "GroupLens: applying collaborative filtering to usenet news," Communications of the ACM (40) 3, 1997.

(9) T. Tokunaga and M. Iwayama, "Text categorization based on weighted inverse document frequency," IPSJ SIG Report. NL100 (5), 1994.

(10) R. Cooley, et al., "Data Preparation for Mining World Wide Web Browsing Patterns," Knowledge and Information Systems, Vol. 1-1, 1999.

(11) B. Mobasher, et al., "Automatic Personalization on Web Usage Mining," Technical Report TR99-010, Department of Computer Science, Depaul University, 1999.

(12) G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar, "Multilevel hypergraph partitioning: Application in VLSI domain," In Proceeings ACM/IEEE Design Automation Conference, 1997.

(13) J. Srivastava, R. Cooley, M. Deshpande, P-T. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," SIGKDD Explorations, (1) 2, 2000.

(14) Sarwar, B. et al., "Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System," Proc. ACM CSCW 98, pp. 345-354, 1998.

(15) Yang, Y., and Liu, X. "A Re-examination of Text Categorization Methods." In Proceedings of ACM SIGIR'99 conference, pp. 42-49, 1999.

(16) 정영미, 정보검색론, 구미무역 출판부, 1993.

(17) 전미선, 박세영, "상호 정보를 이용한 어의 모호성 해소에 관한 연구,"제6회 한글 및 한국어 정보처리 학술발표 논문집, pp. 369-373, 1994.

(18) 박영규, “연관 웹 문서 분류와 브라우징 순차 패턴을 이용한 동적 링킹 시스템,” 인하대학교 대학원 공학 석사 학위 논문, 2000.