

웹 포털 이용자 로그 데이터에 기반한 개인화 검색 서비스 모형의 설계 및 평가*

Design and Evaluation of a Personalized Search Service Model Based on Web Portal User Activities

이 소 영 (Soyoung Lee)**

정 영 미 (Young-Mee Chung)***

초 록

이 연구에서는 한국형 포털에 적합한 커뮤니티 기반 개인화 검색 서비스 모형을 제안하였다. 개인화 검색 서비스 모형은 이용자의 관심 주제를 파악하는 과정과 이를 반영한 검색 결과 재순위화 및 관련 주제 카테고리화 질의어 추천 과정으로 구성된다. 개인화 검색 모형의 유용성을 검증하기 위한 실험에서는 포털 사이트 다음에서 12일간 수집한 이용자 로그 데이터를 사용하였다. 실험 결과 개별 이용자의 주제 카테고리 선정에 사용한 카페 활동성 분석과 신지식 활동성 분석 데이터는 매우 유용한 것으로 나타났으며, 개인화 검색 결과와 추천 서비스에 대한 만족도도 비교적 높게 나타났다.

ABSTRACT

This study proposes an expanded model of personalized search service based on community activities on a Korean Web portal. The model is composed of defining subject categories of users, providing personalized search results, and recommending additional subject categories and queries. Several experiments were performed to verify the feasibility and effectiveness of the proposed model. It was found that users' activities on community services provide valuable data for identifying their interests, and the personalized search service increases users' satisfaction.

키워드: 웹 포털 검색 서비스, 개인화 검색, 이용자 로그 분석, 추천 서비스

Web portal search service, personalized service, log analysis,
recommendation service

1. 서론

야후(yahoo.com)에서부터 본격적으로 시작된 웹 포털 서비스는 정보의 바다라 불리는 인터넷의 활성화와 더불어 비약적으로 발전하여 왔다. 초창기 포털 서비스는 다양한 콘텐츠를 초기 화면에 주제별로 나열하는 방식에 의해 관문 역할을

수행하였으나, 검색 기술이 발전하면서 검색 서비스 위주로 변화하고 있다.

이용자의 정보요구 및 성향이 다양해지고 검색 서비스에 대한 기대 수준이 높아짐에 따라 웹 포털 검색 서비스에서도 개인화와 관련된 다양한 시도가 이루어지고 있다. 이러한 시도는 크게 다음과 같이 네 가지 흐름으로 구분된다.

첫 번째 흐름은 가장 간단한 유형으로서 이용자의 속성 정보에 근거한 개인화이다 (Bonnet 2001). 개인화 서비스의 초기에는 이용자가 입력한 프로파일에 기반하여 맞춤형 검색을 제공하였는데, 이 방식은 대부분 도입 이후 검색 로그 마이닝 기반의 이용자 속성 분석에 의하여 정확도를 보완하고 있다.

두 번째 흐름은 이용자의 행동 정보에 근거하는 개인화로서 검색 이력(history) 위주의 다양한 기능을 제공하는 서비스이다. 2004년 10월 야후가 'My Search' 시범 서비스를 선보이면서 개인화 검색이 검색업계의 화두가 되었다. 야후의 My Search는 이용자가 입력했던 질의어 등의 검색 이력을 제공하고, 검색 결과를 저장하고 편집할 수 있는 기능 위주로 만들어졌다. 구글은 2005년 가을 이용자가 원하는 검색 이력을 포함한 서비스와 주제로 첫페이지를 구성하는 'Personalized Home'이라는 시범 서비스를 선보였으나, 이용률이 저조하여 현재는 제공되지 않고 있다. 이러한 검색 이력 중심의 개인화 서비스는 야후가 2005년 12월 Eurekster(eurekster.com)의 'Social Network Search'라는 사회 연결망 개념을 도입함으로써 개인의 검색 결과를 다른 사람과 공유할 수 있는 모형으로 발전되었다.

세 번째 흐름은 이용자의 자산 데이터를 대상으로 하는 개인화이다. 단순하게 개인 컴퓨터의 데이터를 효율적으로 찾아주는 기존의 소규모 솔루션 위주의 데스크 탑 검색 서비스는, 2004년 말부터 마이크로소프트와 구글 등 초대형 서비스가 주도하면서 전기를 맞았다.

네 번째 흐름은 2003년부터 폭발적인 인기를 얻고 있는 사회 연결망 개념을 도입한 서비스이며, 기본적인 개념은 'my page'에서 저장하고 편집한 정보를 다른 사람과 공유하는 것이다. 이 모형의 문제점은 활발한 공유가 가능한 네트워크를 구축하기가 너무 어렵다는 것이다. 특히 웹 검색 포털에서 자발적으로 검색 결과를 공유하고자 하는 요구가 있는 이용자들이 많지 않아서 아직 성공적인 서비스를 확립한 예는 없다.

이상과 같은 기존의 개인화 서비스 흐름에서 간과되었던 요소로 커뮤니티 서비스가

있다. 한국 웹 포털은 강력한 커뮤니티 기반에 힘입어 서비스가 발전되어 왔기 때문에, 커뮤니티와 연계된 개인화 서비스를 검토할 필요가 있다. 네이버(naver.com)의 지식검색이나 다음(daum.net)의 카페와 연계된 검색 서비스인 카페검색과 신지식 서비스, 그리고 다양한 블로그 검색 서비스 등은 커뮤니티와 연계하여 성공한 대표적인 서비스이다. 이는 실생활에서 정보요구를 접근 가능한 ‘사람’에게 물어보듯이, 웹 상에서 검색 가능한 정보자원의 범위를 공식적으로 출판되지 않은 사람들의 인지 공간까지 넓혔다는 의의가 있다. 커뮤니티 서비스는 이용자가 관심 있는 커뮤니티를 검색하고, 주기적으로 방문하여 정보를 소비하고 생산하기 때문에, 커뮤니티 이용 행태를 분석하여 이용자의 관심사를 유의미하게 파악할 수 있다. 또한 규모 면에서도 커뮤니티 내에서 공유되는 정보의 범위가 이용자의 일반적인 정보요구를 수용할 수 있을 만큼 포괄적이라 할 수 있다.

이와 같은 한국 웹 포털의 특징인 광범위하고 상세한 커뮤니티에 기반한 확장형 개인화 검색 서비스 모형을 개발하는 것이 이 연구의 목적이다. 구체적으로 이용자의 관심주제를 효율적으로 추출하기 위한 이용자 관심 주제 카테고리 선정 모형과 질의어 주제 카테고리 선정 모형을 제안하고, 이 모형의 유용성을 실험을 통해 평가한다. 또한, 유사한 관심 주제를 공유하는 이용자들의 정보를 협업 필터링 방식에 의해 추출하여 이용자의 정보요구와 관심사를 확장시킬 수 있는 방법을 제안한다.

2. 개인화 검색 서비스

2.1 개인화 검색 서비스의 유형

일반적으로 개인화(personalization)라는 용어는 이용자 정보요구에 부합되는 콘텐츠를 제공한다는 의미로 광범위하게 사용된다(Shahabi & Chen 2005). Riecken(2000)은 “의미있는 1 대1 관계를 형성하여 이용자의 서비스에 대한 충성도(service royalty)를 제공하는 것”으로 규정하였는데, 특히 이용자의 충성도 부분은 특히 포털 서비스에서 가장 중심적으로 추구하는 개념이다. 개인화와 관련하여 고객화(customization)와 추천(recommendation)의 개념이 있다. 고객화는 이용자의 선택 및 수정에 기반한 이용자의 선택권을 가장 중요하게 보며(황용석,

이흥철 2000), 추천은 이용자의 선호에 의해서 대규모의 대상(items)들로부터 유용한 대상을 제안하는 개념이다(Johansson 2004).

Perugini(2004)는 개인화가 정보를 이용자에게 전달하는 데 필요한 절차와 기술로 구성되며, 이는 정보, 구조, 표현으로 정의된다고 보았다. 김영환 (2002)은 이용자가 “선호하거나 필요로 하는” 콘텐츠를 “선호하거나 필요로 하는” 방식으로 제공하는 것으로 정의하였다. 그는 개인화의 대상을 인터페이스, 커뮤니케이션, 상품 등으로 나누어 보았다.

개인화 검색의 유형은 개인을 정의하는 방법에 따라서 구분하는 방식과, 어떤 내용을 개인화하는가에 따라서 구분하는 방식으로 나누어 살펴볼 수 있다.

첫 번째 유형 구분 방식은 ‘개인’을 정의하는 방법에 따라서 서비스 유형을 구분한다(Bonnet 2001). 첫째, 개인의 프로파일을 생성하는 방법으로서 이용자가 적극적으로 프로파일의 사항을 입력하면, 이를 이용하여 기본 프로파일을 작성하는 방식이다. 둘째, 서비스 이용행태를 기반으로 하는 방식으로서 클릭의 흐름을 분석하여 이용자의 프로파일을 정의하는데, 보통 웹 이용 마이닝 시스템을 통하여 작성한다. 셋째, 이용자의 사회화 프로파일(social profile) 기반 방식으로서 이용자가 누구와 관계를 맺고 있는가를 추적하여 협업 필터링을 통하여 개인을 정의한다.

두 번째는 어떤 내용을 개인화하여 검색 결과를 제공하는지에 따른 구분 방식이다. 링크정보를 이용하거나, 질의를 확장하거나, 결과를 재순위화하거나, 메타 검색, 혹은 도메인별 검색 등이 있다. Jeh and Widom(2003)은 이용자가 즐겨 찾는 페이지에서 링크되거나, 해당 페이지가 링크한 페이지에 더 많은 가중치를 두어 검색랭킹에 반영하는 형태를 연구하였다. 또한 개인에게 맞추어진 방법으로 질의를 보정하거나 확장하는 방법도 연구되었다. Liu, Yu, and Meng (2002)은 이용자의 프로파일에 기반해 입력된 질의를 이용자 관심 주제로 매핑시켜 검색 대상을 한정시켰다. 또 Websfter 프로젝트에서는 이용자의 검색 의도를 입력받은 후, 이를 해당 이용자의 검색 분류체계(search taxonomy) 안에서 가장 적절한 질의어로 변환시키는 방법을 연구하였다(Scime and Kerschberg 2000). Tanudjaja and Mui(2002)는 그래프 모형에 기반한 신뢰도에 의해 랭킹을 조정하여 상위 결과만을 제공하는 방식으로 검색 결과를 필터링하는 Persona를 개발하였다. Persona는 이용자의 피드백을 통해 검색결과를 정제하는 모형도 포함한다.

또한 개인화된 메타검색 엔진들도 개발되었는데, Inquirus2는 사용자가 원하는 주제 분야를 선정하고, 검색엔진을 선택할 때 해당 주제의 내용을 이용하였다(Glover et al. 2000). Zhu 등(2001)은 사용자가 'good'이나 'bad'로 평가한 사항을 기반으로 메타 검색엔진의 검색결과를 하나로 보여주는 방식을 연구하였다.

2.2 개인화 검색 서비스 구현의 문제점

웹 포털에서 개인화 검색 서비스를 제공하기에는 여러 가지 현실적인 어려움이 따른다. 첫째는 이용자의 관심 주제 추출과 관련된 문제이다. 이용자가 직접 관심 주제를 입력하는 것이 개인화 서비스의 유의미하고 안정적인 출발선일 수 있다. 그러나 이용자가 직접 관심 주제를 입력하는 작업이 매우 번거롭기 때문에 참여율이 저조할 수 밖에 없다. 또한 이용자의 관심주제가 시간에 따라 변하는데, 그때마다 이용자가 관심주제를 수정할 것으로 기대하기 어렵다. 이와 같은 방식은 변화되는 이용자의 관심 주제를 효율적으로 반영하지 못한다. 일반적으로 적극적인 이용자 그룹은 전체 이용자의 0.5% 미만이고, 이 0.5%를 전체로 확대하기에는 샘플링의 보편성을 보장할 수 없기 때문에 이용자의 불편을 최소화 하면서, 동시에 이용자에 대한 정보를 풍부하게 수집해야 하는 딜레마가 내재되어 있다.

둘째는 로그 데이터가 가지는 단편성이다. 로그 데이터는 이용자의 서비스 사용에 대한 완전한 맥락 정보를 갖지 않은 단편적인 정보를 저장한다. 로그 데이터를 이용해 이용자의 관심 주제를 추출할 경우, 광범위한 이용자의 관심 주제를 포괄적이고 입체적으로 반영하기 힘들다. 다시 말해 이용자의 의도나 해석과 같은 정량적으로 측정하기 힘든 부분에 대한 파악이 실질적으로 불가능하다는 것이다.

셋째는 처리해야 하는 방대한 데이터 양이다. 분석 기간이나 범위 등에서 지나치게 포괄적인 분석을 시도하거나 분석 모델링이 복잡할 경우에는 처리 속도 등의 문제로 적기에 이용자의 관심 주제를 추출할 수 없다. 따라서 적절한 수준에서 분석 범위와 임계치를 설정하는 것이 필수적이다.

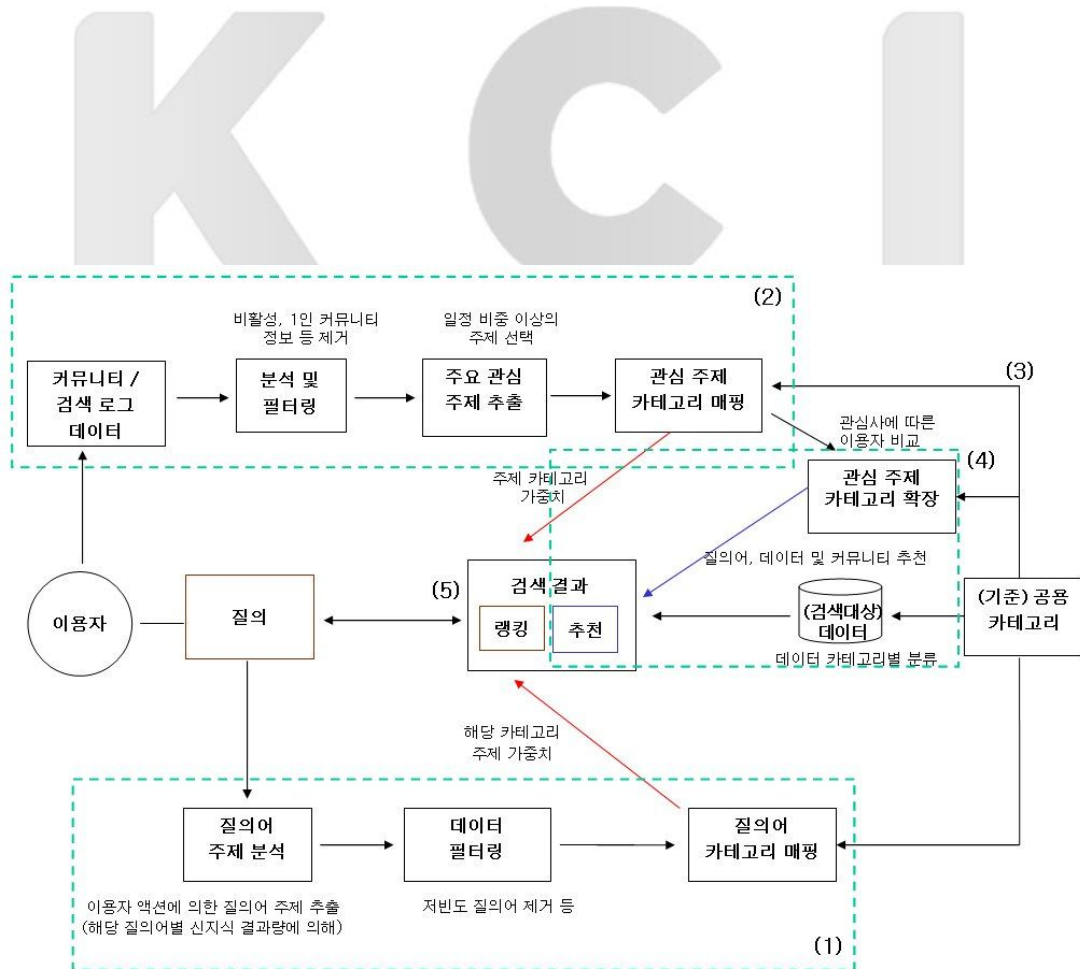
위와 같은 제한점을 극복하기 위해서는 이용자가 스스로 관심사를 등록하는 것과 같은 적극적인 참여를 최소화하는 방향으로 서비스를 구현하는 것이 바람직하다. 본 연구에서 제안한 개인화 검색 모형에서는 포털 사이트 다음(daum.net)에서 이용자의

관심이 적극적으로 반영된 카페 서비스와 검색 서비스의 이용자 로그 데이터를 사용하여 이용자의 관심주제를 자동으로 찾아내도록 하였다. 카페 서비스에서는 이용자가 관심 있는 카페에 가입하여 주기적으로 방문하고 글을 읽거나 쓰는 적극적인 이용패턴을 보이며, 검색 서비스도 질의어를 입력하고 검색결과를 클릭하는 적극적인 이용패턴을 보이기 때문이다.

3. 개인화 검색 모형

3.1 모형 개요

개인화 검색 시스템은 이용자 프로파일 데이터베이스, 검색 솔루션, 그리



<그림 1> 개인화 검색 모형 개요

고 개인화 서비스의 세 요소로 구성된다. 이용자 프로파일 데이터베이스는 수집한 이용자의 로그 데이터를 주간이나 격주간 등의 일정한 기간 단위로 일괄처리하여 생성한다. 프로파일 데이터베이스에는 이용자의 관심 주제 카테고리 및 관련성 점수, 질의어 주제 카테고리 및 관련성 점수 등이 저장되고, 이를 기반으로 개인화 서비스를 제공하게 된다. 검색 솔루션에는 신지식 검색과 웹사이트 검색이 있다.

<그림 1>은 개인화 검색 과정을 보여주며, 이를 단계별로 기술하면 다음과 같다.

(1) 전체 이용자 질의어의 주제 분석 및 카테고리 매핑

이 단계에서는 일정 기간 입력된 모든 질의어의 주제 분석을 시도한다. 질의어를 신지식 서비스에서 검색하여, 카테고리별로 그룹핑한 검색결과 건수에 따라 주제 카테고리를 설정하였다. 이용자가 신지식 서비스에 질문을 등록할 때 주제를 잘못 선택하는 오류가 있을 수 있으나, 대규모 이용자를 통해 관심있는 주제에서 질문 및 답변이 생성되므로 오류의 비율은 그다지 크지 않다. 주제 카테고리는 기본적으로 신지식 서비스에 이용되는 900여 개의 카테고리 구조를 기준 카테고리로 이용한다

(2) 이용자의 커뮤니티/검색 행태에 대한 로그 데이터 분석

커뮤니티 서비스와 검색 서비스의 로그를 분석하여 이용자의 관심 주제를 추출한다. 커뮤니티 서비스는 다음 카페 서비스(cafe.daum.net)를 의미한다. 포털 사이트 다음(daum.net)에서 제공하는 회원제 커뮤니티 서비스인 카페 서비스는 모든 분야를 포괄하며 약 600만개 정도가 개설되어 있다. 2005년 수행된 한 설문 조사에서 카페를 이용하는 첫 번째 이유가 정보를 얻기 위해서라고 응답할 만큼, 단순한 친목 유지의 이유뿐만 아니라, 정보 습득을 목적으로 카페 서비스를 이용하고 있다.

(3) 이용자 관심사의 기준 카테고리 매핑

이용자의 가입 카페 및 활동 정보를 활용하여 이용자 개인의 관심 주제 카테고리를 선정한다. 이용자가 어떤 카페에서 어느 정도의 활동을 하는가를 분석하고, 해당 카페가 기준 카테고리의 어떤 주제에 어느 정도의 점수로 매핑되는가를 판단한다.

이용자의 검색 활동 행태는 검색 질의어를 분석하여 파악한다. 이용자 개인별로 검색 질의어의 빈도를 계산하여, 질의어에 매핑되는 주제 카테고리의 점수를 산출한다. 신지식 서비스는 이용자가 생산하는 질문, 답변 데이터를 추출하여, 주제 카테고리의 점수를 산출한다. 이러한 방식으로 얻어진 커뮤니티 활동과 검색 활동에 따라 매핑된 주제 카테고리의 점수를 합산하여 이용자 별로 최종 관심 주제 카테고리를 선정한다.

(4) 이용자 관심 주제 카테고리의 확장

이용자의 관심 주제 프로파일을 작성하고, 유사한 주제 프로파일을 공유하는 사람들의 관심 주제를 추천한다. 즉, 동일 그룹의 구성원들이 공통으로 가지고 있는 관심사를 이용자의 미래의 정보 이용에 도움이 될 수 있는 주제 카테고리로 추천한다. 기본적으로 이용자에게 관심이 있을 만한 주제를 추천하고, 이 주제에 관련된 질의어나 인기 데이터도 제공한다.

(5) 검색결과물의 재순위화

선정된 이용자의 관심 주제 카테고리에 따라 검색결과에 가중치를 부여하여 개인화된 검색결과를 순위화한다. 또한 질의어에도 선정된 카테고리에 따라 가중치를 부여한다.

3.2 질의어 주제 카테고리 매핑 과정

다음 검색 서비스(search.daum.net)에서 일정기간 입력된 각 질의어가 어떤 주제 카테고리와 밀접한 관련이 있는지를 분석하여 가장 적절한 주제 카테고리로 매핑하는 과정은 다음과 같다.

먼저 각 질의어에 따라 검색된 결과를 해당되는 주제 카테고리로 매칭하여, 질의어별로 대응되는 각 주제 카테고리의 점수(Qcate_score)를 산출한다. 의미있는 수준의 주제 카테고리로 매핑하기 위하여, 카테고리 점수가 해당 질의어에 대한 카테고리 점수 평균 이상인 카테고리만을 질의어에 대응되는 카테고리로 설정하였다. 마지막으로 질의어별 주제 카테고리 점수를 질의어간 수평적 비교를 위하여 정규화하여 “정규화 질의어-카테고리 점수”(NQcate_score)를 산출하고 이 점수를

질의어에 대응되는 주제 카테고리 매핑에 사용한다.

3.3 이용자 주제 카테고리 선정 과정

3.3.1 카페 서비스 이용 분석

(1) 이용자별 카페 활동도 산출 및 카페 선정

이용자의 카페 서비스 이용 분석을 위하여, 이용자별 가입 카페 리스트를 확보한 다음 이용자가 가입한 카페에서 얼마나 활발하게 활동하고 있는가를 파악하기 위해 각 이용자의 카페별 활동도를 산출한다.

각 이용자의 카페별 활동도인 “이용자 카페 활동도(Ucafe_score)”를 측정하는 요소는 방문 횟수(Ucafe_visit_freq), 카페 내 게시(쓰기) 수(Ucafe_write_freq), 클릭(읽기) 수(Ucafe_read_freq) 등이다. 기간 중 방문 횟수(Ucafe_visit_freq)는 카페를 클릭하여 개별 카페의 탑 페이지를 오픈한 횟수이다. 이에 반하여, 특정 게시글을 읽는 행위(Ucafe_read_freq)는 최소 2회 이상의 클릭이 선행되어야 하고, 최종 콘텐츠에 관심이 있는 것을 반영하기 때문에 가중치 2를 부여하였다. 글을 생성하는 행위(Ucafe_write_freq)는 적극적인 행위이기 때문에, 클릭 행위의 2배에 해당하는 가중치 4를 부여하였다. 최종적인 산출 공식은 다음과 같다.

$$Ucafe_score = a * Ucafe_visit_freq + b * Ucafe_write_freq + c * Ucafe_read_freq \quad (\text{단, } a=1, b=4, c=2)$$

위와 같이 산출된 “이용자 카페 활동도”(Ucafe_score)에 따라, 이용자가 핵심적으로 이용하는 카페를 선별하였다. 1인당 평균 카페 수가 18~20개 정도인 점을 고려하면, 가입하여 활동하는 모든 카페를 관심분야로 설정하는 것은 잡음을 유발할 수 있다. 따라서, 개인의 “개별 카페 활동도”가 개인의 활동도 평균값 이상이 되는 카페로 관심 카페를 제한하였다. 위와 같이 산출한 “이용자 카페 활동도(Ucafe_score)”를 카페 활동도 총합으로 나누어 “정규화 이용자 카페 활동도”(NUcafe_score)를 산출한다.

(2) 카페별 주제 카테고리 매핑

이 단계에서는 앞에서 이용자별로 선정한 카페들을 주제 카테고리화 매핑하는 작업을 수행한다. 이를 위해 해당 카페에 설정된 지식게시판의 주제와 카페 검색에서 해당 카페가 클릭된 로그를 분석한다.

신지식 게시판 분석을 통한 카페의 주제 카테고리 매핑 방식은 다음과 같다. 각 카페에 대한 “지식게시판-카테고리 점수(Ckin_cate_score)”는 카페의 해당 지식게시판 활동도로 측정하였다. 개별 지식게시판 활동도는 카페 메뉴 안의 지식게시판에 게시된 질문 수와 신지식 서비스 상에서 게시한 질문 수를 합한 값(Ckin_cate_q_freq)과 지식게시판과 신지식 서비스 상의 답변 수를 합한 값(Ckin_cate_a_freq)을 각각 산출한 다음 질문의 2배의 가중치를 답변에 부여한 다음 합계를 내어 산출한다.

$$Ckin_cate_score = a * Ckin_cate_q_freq + b * Ckin_cate_a_freq$$

(단, $a=1$, $b=2$)

이렇게 각 지식게시판의 주제 카테고리에 대한 점수를 산출하고, 평균값 이상의 지식게시판 주제 카테고리 점수를 보유한 게시판들만을 분석 대상으로 선정하였다. 해당 지식게시판에 대한 주제 카테고리 점수(Ckin_cate_score)를 해당 카페 지식게시판 총 점수로 나누어 “정규화 지식게시판-카테고리 점수”(NCkin_cate_score)를 산출한다.

카페 검색의 클릭로그 분석을 통한 카페의 주제 카테고리 선정 방식은 다음과 같다. 질의의 결과로 카페를 클릭하였을 경우, 해당 질의어와 카페를 연결하여 로그 데이터를 수집한다. 여기에서 질의어에 기분석된 주제 카테고리를 클릭 카테고리로 매핑하고, 해당 카페 클릭 횟수를 이용하여 “클릭로그-카테고리 점수”(Csearch_cate_score)를 산출한다. 카페 클릭로그의 경우, 한 카페가 여러 개의 검색 질의어에 반응할 수 있고, 또한 한 개의 검색 질의어가 여러 개의 카테고리에 매핑될 수 있기 때문에, 각 카페는 한 개의 카테고리에 대해서 여러 개의 값을 가질 수 있다. 이 경우 최대값을 선택하여 매핑에 사용하였다.

일정 수준 이상의 카테고리 점수를 가진 카테고리로 제한하기 위하여, 카페의 “클릭로그-카테고리 점수” (Csearch_cate_score)가 평균치 이상인 것만 선정한 다음 “정규화 클릭로그-카테고리 점수” (NCsearch_cate_score)를 산출한다.

마지막으로 카페의 지식게시판과 검색의 클릭로그 등 두가지 데이터를 근거로 산출된 “정규화 지식게시판-카테고리 점수”(NCkin_cate_score)와 카페의 “정규화 클릭로그-카테고리 점수”(NCsearch_cate_score)를 합산하여 “정규화 카페-카테고리 점수”(NCcate_score)를 산출한다.

(3) 이용자별 카페기반 주제 카테고리 점수 산출

앞에서 산출된 “정규화 이용자 카페 활동도”(NUcafe_score)”와 “정규화 카페-카테고리 점수”(NCcate_score)를 곱하여 “이용자 카페기반 카테고리 점수”(NUcafe_cate_score)를 산출한다. 이 점수는 최종적으로 이용자별 주제 카테고리 선정에 사용된다.

3.3.2 신지식 서비스 이용 분석

이용자 주제 카테고리 선정에는 카페 서비스와 함께 신지식 서비스의 이용 분석이 사용된다. 즉 이용자가 신지식 서비스의 어떤 카테고리에서 질의하거나 답변하는지를 분석하여 관심 주제를 추출하며, 데이터로는 이용자별 신지식 이용 로그를 취합하여 사용한다.

먼저 각 이용자의 “이용자 신지식 카테고리 활동도”(Ukin_cate_score)는 다음과 같이 산출한다. 활동을 측정하는 요소로는 질문 수(Ukin_cate_q_freq), 답변 수(Ukin_cate_a_freq)를 사용하며, 답변의 가중치를 질의의 2배로 조정하였다.

$$Ukin_cate_score = a * Ukin_cate_q_freq + b * Ukin_cate_a_freq$$

(단, $a=1$, $b=2$)

“이용자 신지식 카테고리 활동도” (Ukin_cate_score)를 해당 이용자의 신지식 총 점수로 나누어 정규화한 “정규화 이용자 신지식 카테고리 활동도”(NUkin_cate_

score)를 이용자 주제 카테고리 선정에 사용한다.

3.3.3 질의어에 대한 검색 결과 클릭 데이터 분석

이용자가 검색 서비스를 통해 입력한 검색 질의어에 대한 검색 항목 중 클릭된 정보를 이용하여 이용자가 입력한 질의어에 대한 카테고리 점수를 산출하여 이용자 주제 카테고리 선정에 이용한다. 이때 이용자가 입력한 질의어가 많을 경우 1개 이상의 카테고리 점수가 계산될 수 있다. 이 경우 매칭된 카테고리 점수 중 최대값을 적용한다.

이용자의 질의어 입력 빈도에 따른 “이용자 질의어 점수”(Uqry_score)와 질의어 주제 카테고리 매핑에서 도출된 해당 질의어의 “정규화 질의어-카테고리 점수”(NQcate_score)를 곱하여 “이용자 질의어-카테고리 점수”(Uqry_cate_score)를 산출한다.

$$Uqry_cate_score = Uqry_score * NQcate_score$$

“이용자 질의어-카테고리 점수” (Uqry_cate_score)를 총 점수로 나눈 “정규화 이용자 질의어-카테고리 점수”(NUqry_cate_score)를 이용자 주제 카테고리 최종 선정 공식과 이용자 그룹핑에서 사용한다.

3.3.4 이용자 주제 카테고리 최종 선정

위와 같이 분석한 카페, 신지식 서비스, 검색 서비스 데이터를 취합하여 최종적인 이용자 주제 카테고리를 선정한다. 아래 공식에서 “이용자-카테고리 점수”(Ucate_score)는 “정규화 이용자 카페기반 카테고리 점수” (NUcafe_cate_score)와 “정규화 이용자 신지식 카테고리 활동도”(NUkin_cate_score), “정규화 이용자 질의어-카테고리 점수 (NUqry_cate_score)”를 각각의 가중치를 부여하여 더한 값이 된다.

$$U_{cate_score} = a * (NU_{cafe_cate_score}) + b * NU_{kin_cate_score} + c * NU_{qry_cate_score}$$

위의 공식에서 사이트의 특성이 반영되어야 하지만 포털 사이트 다음(daum.net)과 같은 경우는 카페 서비스가 검색 서비스에 비해 월등히 높은 이용도와 충성도를 보인다. 본 연구에서는 분석한 데이터의 양을 고려하여 a와 b의 비율을 11.16:1로 설정하였다. 질의어 로그 데이터의 경우 로그인 이전에 입력한 질의어는 사용하지 않았으며, 필터링 임계치를 5회로 설정하였기 때문에 필터링된 질의어가 많았다. 본 연구에서는 데이터 자체에서 필터링된 비율을 고려하여 상대적으로 가중치 값을 적게 설정하여야 할 것으로 보고 b와 c의 비율을 1: 0.3으로 정하였다.

위의 공식에 의해 산출된 주제 카테고리의 카테고리 점수(Ucate_score)가 평균 카테고리 점수 이상인 카테고리를 이용자의 관심을 반영하는 최종 주제 카테고리로서 선정한다.

3.4 이용자 그룹핑에 의한 주제 카테고리 및 질의어 추천

이용자의 주제 카테고리로서 확정된 카테고리는 이용자가 “현재” 관심을 보이고 있는 활성화된 주제라고 할 수 있다. 이러한 관심 주제를 검색 결과에 반영함으로써 이용자의 정보요구에 대한 만족도를 향상시킬 수 있다. 반면 추천은 이용자의 관심 주제를 협업 필터링을 통해 “미래”에 관심을 가질 수 있는 범위까지 확장시키기 위한 장치라고 할 수 있다.

추천 서비스를 통해 추천 주제 카테고리의 제시뿐만 아니라, 추천 카테고리에 속하는 인기 질의어, 인기있거나 최신의 데이터(신지식, 뉴스, 웹사이트 등), 관련된 카페 리스트 등이 제공될 수 있다.

본 실험에서 관심 카테고리를 추천하기 위해 “외부 확장” 방식과 “내부 부활” 방식의 두 가지 방식을 사용하였다. 외부 확장 방식은 한 이용자의 관심 주제 카테고리 리스트를 다른 이용자와 비교하여, 각 이용자의 리스트 크기(관심 카테고리 개수)와 내용(실제 주제 카테고리)이 75% 이상 일치하는 경우, 다른 이용자의 나머지 카테고리를 관심 카테고리로서 추천한다. 즉, 크기에 의한 1차 이용자 그룹핑이

완료되면, 그룹내의 모든 이용자를 대상으로 주제 카테고리 리스트 내용을 비교하여 2차 사용자 그룹핑이 이루어진다. 각 이용자의 주제 카테고리 리스트는 벡터 형태로 표현하고, 사용자 벡터간 유사도 계산을 위하여 코사인 유사계수를 사용하였다. 특정 이용자와 같은 사용자 그룹에 속한 주제 카테고리 리스트를 생성한 다음 이 리스트에서 높은 출현 빈도를 기록한 주제 카테고리를 이용자에게 추천한다.

두 번째 방식인 내부 부활 방식은 한 이용자의 관심 주제 카테고리에서 관심 주제로 인식은 되었으나, 임계치로 설정된 평균 점수를 넘지 못해 관심 카테고리로 확정되지 못한 주제 카테고리를 후보그룹으로 형성한다. 이 이용자의 주제 카테고리 리스트의 크기와 내용이 유사한 주제 카테고리를 가진 다른 이용자와 비교하여 더 많은 사람들과 공유가 많이 되는 순으로 카테고리를 추천하는 방식이다.

질의어 추천은 앞에서 기술한 방법에 의해 추천된 주제 카테고리에서 높은 “정규화 질의어-카테고리 점수” (NQcate_score)를 얻은 n개의 질의어를 추천한다. 복수의 카테고리가 추천되는 경우에는 첫째, 추천 카테고리 선정된 모든 카테고리마다 최상위 1개씩의 질의어를 추천하거나, 둘째, 추천 카테고리 자체의 가중치에 비례하여 추천하는 해당 카테고리의 질의어 수를 변화시킨다. 질의어 추천과 유사한 방식으로 정규화 카페 카테고리 점수(NCcate_score)를 이용하여 카페를 추천한다.

4. 실험 결과 및 분석

4.1 실험 설계

실험을 위한 플랫폼은 리눅스 기반의 PC 서버에 구성하였고, 대부분의 데이터 처리는 ORACLE 9e 버전을 이용하였다.

기본적인 실험과정은 다음과 같다. 첫째, 질의어와 이용자를 대상으로 대응되는 주제 카테고리를 선정한다. 둘째, 이용자의 주제 카테고리에 따라 질의어에 주제 가중치를 부여하여 검색 결과를 순위화한다. 셋째, 사용자 그룹핑에 의해 새로운 주제 카테고리를 추천한다. 넷째, 새로 추천된 주제 카테고리과 관련된 핵심적인 질의어를 추천한다. 구체적인 실험 과정은 <그림 2>와 같다.

<그림 2> 데이터 수집 과정

실험을 위한 이용자의 로그 데이터는 2006년 4월 10일부터 4월 20일까지 12일간 수집된 데이터로서, 검색 로그, 카페 주제 카테고리 매핑 로그, 신지식 서비스 로그 데이터 등이다. 일반적으로 실제 서비스 환경에서는 대용량 데이터의 처리를 위하여 1주일 정도의 기간을 기준으로 데이터를 처리하는 경향이 있다. 본 연구에서는 12일간 총 19만여 명의 로그 데이터를 확보하였다. 실제 실험 서비스 이용 목적의 로그 데이터를 확보하였다. <표 1>은 구체적인 로그 데이터 종류와 이용 목적 등을 보여 준다.

<표 1> 이용자 로그 데이터 요약

종류	역할 및 목적	건수
검색 이용자 로그	<ul style="list-style-type: none"> 이용자 추출 검색 결과 순위화 카페 이용정보를 이용하여 이용자의 관심 주제 카테고리 추출 	약 19만명
카페 데이터	<ul style="list-style-type: none"> My카페 리스트, 이용자 관심 카페 활동도 카페 정보(회원 수, 키워드 등) 	약 340만개
카페검색 클릭로그	<ul style="list-style-type: none"> 카페 회색과 평가판 추천 결과 평가 	약 550만개
신지식 데이터	<ul style="list-style-type: none"> 각 카페의 주제 카테고리 매핑 검색 서비스 상의 이용자 주제 카테고리 	약 1,300만 건
검색 질의어 로그	<ul style="list-style-type: none"> 신지식 서비스 활동도 산출 검색 질의어 및 해당 클릭 정보 추출 	2,200만건
		21만건

분석에 이용한 검색 데이터는 실험 기간 중 취합된 모든 질의어 데이터를 대상으로 하나, 일 5회 이하 입력된 질의어를 제외하였다. 저빈도 질의어의 경우에는 오타 등에 의한 예외적 상황에서 발생할 가능성이 많아 질의어의 보편적 신뢰성을 보장하기 힘들다. 또한 빈도 수가 기준치 이하일 경우, 입력한 이용자도 대개 소수로 한정되어, 개인의 취향 등 개별적인 특성이 강하게 반영되므로 보편적인 데이터로 보기 어려운

측면이 있다.

기간 중 취합된 모든 질의어를 신지식 데이터베이스에 질의하여 검색결과가 10건 이상되는 카테고리를 1차적으로 선정하여 카테고리 매핑을 시도하였다. 주제 카테고리 매핑에는 신지식 데이터베이스 분류에 사용된 기준 카테고리를 이용하였다.

본 연구의 실제 서비스에서 이루어지는 이용자 피드백에 따른 데이터의 주제 카테고리 부여 작업은 불가능하기 때문에 검색 대상 데이터베이스를 이미 구조화되어 있는 신지식 서비스 데이터와 디렉토리에서 구조화된 웹사이트 데이터로 제한하였다. 신지식 서비스 데이터와 웹사이트 데이터는 각각 2,200만 건과 28만 건이다.

실험 결과 평가는 실제로 시스템을 구축하여 서비스에 반영한 후 이용자의 피드백 결과를 수집하는 것이 바람직하지만 본 실험에서는 실제 서비스가 구현되지 않았기 때문에, 20명의 표본 이용자에 대하여 데모 시스템을 구성한 후 다음의 네 가지 평가 문항의 설문에 대한 응답을 분석하는 것으로 대체하였다.

1. 관심사로 선정된 카테고리가 적절한가?
2. 관심사로 선정된 질의어가 적절한가?
3. 관심사에 따라 검색결과가 적절하게 순위화 되었는가?
4. 추천된 주제 카테고리가 적절한가?

첫 번째 문항은 선정된 주제 카테고리가 적절한가에 대한 것이다. 개인별로 복수개의 주제 카테고리가 선정될 수 있으며, 각 주제 카테고리마다 적절성 여부를 '예', '아니오'로 답변하도록 하였다. 두 번째 문항은 선정된 질의어가 적절한가에 대한 것이다. 이 문항은 첫 번째 문항의 주제 카테고리에 따라 선정된 질의어 추천이기 때문에, 첫 번째 문항의 결과와 연계하여 분석하였다. 질의어를 추천하는 방식은 (1) 카테고리의 가중치의 비율로 질의어 개수를 조절하는 방식과 (2) 선택된 카테고리에 대해서 무조건 일정 개수를 추천하는 방식을 비교하였다.

세 번째 문항은 선정된 주제 카테고리 및 그 가중치에 의해 2번째 문항에서 제시된 질의어의 결과를 새로 순위화한 검색 결과를 이용자에게 제시하고, 순위화의 적절성을 평가하는 것이다. 상위 10개 검색 결과에 대해서 5점 척도에 의해 개별적인 만족도를 측정하였다. 또한 일반 검색 결과와의 비교를 통하여 개인화된 검색 결과의

품질을 검증하였다.

마지막 문항은 협업 필터링 기반에 의해 추천되는 카테고리의 유효성을 검증하기 위한 것이다. 관심 카테고리를 추천하는 두 가지 방식에 따라 추천된 각각의 카테고리에 대해 5점 척도를 통해 만족도를 측정하였다.

4.2 질의어-주제 카테고리 매핑 결과

각 질의어는 최대 3개의 주제 카테고리로 매핑하였으며, 질의어와 주제 카테고리간 적합성은 다음커뮤니케이션 웹사이트 주제 전문가 3인이 5점 척도를 사용하여 평가하였다. 질의어는 상위빈도 질의어 20개(1일 검색빈도 50,000회 이상), 중위빈도 질의어 20개(1일 검색빈도 100회), 하위빈도 질의어 20개(1일 검색빈도 10회)를 대상으로 하였다.

상위빈도 질의어 20개에 대한 질의어와 주제 카테고리 매핑 실험에서 전문가에 의한 카테고리 적합성 평가 결과 20개 질의어에 대한 평균 점수는 4.38로서 비교적 높은 적합성을 보였다. 상위빈도 질의어의 유형을 살펴보면 “싸이월드”와 같은 특정 사이트 이름과 인물명, TV 프로그램명과 같은 이슈성 키워드가 대부분을 차지한다. 실험 결과 적합성이 가장 떨어지는 이슈성 키워드는 시간이 지나면 다른 의미를 지닐 수 밖에 없다.

상위빈도 질의어 중에 주제 카테고리 매핑에서 적합성이 떨어지는 것은 대부분 갑자기 검색 빈도가 급증하는 것들이다. 또한 질의어한 이용자의 관심주제가 커뮤니티나 포털이라고 보기는 어렵다.

중위빈도 질의어 20개에 대한 질의어의 주제 카테고리 매핑에서 전문가에 의한 카테고리 적합성 평가 결과, 20개 질의어에 대한 평균 점수는 4.28로서 비교적 높은 적합성을 보였다. 부적합하게 매핑된 질의어들은 관심 주제와 관련한 질의어라기 보다는 단순히 사이트를 찾고자 하는 의도에서 비롯된 것으로 보인다.

하위빈도 질의어 20개에 대한 주제 카테고리 매핑에서는 적합성 평가 결과 평균 점수는 4.1로 나타났다. 부정확하게 질의어의 경우는 검색 결과 수가 6개 이하로 적어, 잘못 매핑될 가능성이 높을 수 밖에 없었다.

<표 2>는 질의어의 빈도 범위별로 검색 건수와 적합성 점수를 비교한 것이다. 세

빈도 유형의 질의어들의 평균 적합성 점수는 4.25로 높게 나타났으며, 상위빈도 질의어가 하위빈도 질의어에 비해서 주제 카테고리 매핑이 좀 더 정확한 것을 알 수 있다. 하위빈도 질의어가 상위빈도 질의어에 비해서 주제 매핑 적합성이 약간 떨어지는 원인은 검색 건수의 차이에서 비롯된다고 볼 수 있다. 검색 건수와 적합성 점수간 상관계수는 0.9로 매우 상관성이 높게 나타났다.

<표 2> 질의어 빈도에 따른 주제 카테고리 적합성 비교

질의어 유형	검색 건수	적합성 점수
상위빈도 질의어	169	4.38
중위빈도 질의어	47	4.28
하위빈도 질의어	9	4.10

4.2 이용자 관심 주제 카테고리 선정 실험 결과

본 실험에서는 이용자의 관심 주제 카테고리를 선정하는 기준으로 카페 활동, 신지식 활동, 검색 질의어 등 3가지 요소를 사용하였다.

실험기간 동안의 데이터를 기반으로 한 실험 모집단을 이루는 다음 카페, 신지식 서비스, 검색 서비스 사용자들의 특성은 <표 3>과 같다. 현재 활동중인 카페 회원은 무려 700만 명에 이르고 활동중인 신지식 회원은 60만 명이지만 활동중인 검색 이용자는 매우 적었다. 실험 모집단의 규모로 볼 때 카페 활동을 통한 이용자 주제 카테고리가 가장 유효한 기준이며, 검색과 신지식 활동은 보조적인 기준으로 판단하였다.

<표 3> 이용자 집단의 특성

	활동성있는 카페회원 (3회 이상 방문)	일반 카페회원 (3회 미만 방문)	활동성있는 신지식회원 (활동지수20 이상)	일반 신지식 회원 (활동지수20미 만)
활동성있는 검색 서비스 이용자 (일평균 6회이상 검색)	170,355 (2.4%)	6,244 (0.4%)	73,295 (12%)	106,292 (7%)
일반 검색	3,879,177	156,668	378,692	485,910

서비스 이용자 (일평균 5회이하 검색)	(55.6%)	(9.3%)	(62%)	(32%)
검색 서비스 비이용자	2,927,706 (42%)	1,527,821 (90%)	158,806 (26%)	926,266 (61%)
총합	6,977,238 (100%)	1,690,733 (100%)	610,794 (100%)	1,518,469

표본으로 선정한 20명의 이용자에 대해 카페 활동도, 신지식 활동도, 검색 활동도를 모두 고려하여 “이용자-카테고리 점수”(Ucate_score)를 산출하고, 이 점수가 이용자의 평균 주제 카테고리 점수 이상인 카테고리를 이용자의 최종 관심 주제 카테고리로 선정하였다.

실험결과 20명 중에 17명에게는 1개의 관심 주제 카테고리가 선정되었고, 2명에게는 2개, 1명에게는 4개의 주제 카테고리가 선정되었다.

4.3 검색 결과 재순위화 실험 결과

검색 결과의 적합성 평가 실험은 이용자 프로파일이 생성된 20명을 대상으로 하여, 중위빈도 질의어 중에서 무작위로 선정한 질의어 10개씩과 주제와 관련된 질의어 10개씩, 총 400개의 질의어를 가지고 수행하였다. 상위빈도 질의어는 대부분 사이트명이나 연예인명, 특정 이슈와 관련된 특징을 보였으므로, 관심 주제와 관련성이 높은 중위빈도 질의어를 선정하였다.

검색 결과의 적합성 평가는 신지식 검색 결과 상위 3개, 웹사이트 결과 상위 2개로 제한하고, 5점 척도를 사용하여 개인화된 검색과 그렇지 않은 검색 결과를 비교 평가하였다.

개인화 검색 결과의 알고리즘은 기존 서비스에서 이용중인 신지식과 웹사이트의 순위화 알고리즘을 그대로 사용하되, 이용자 프로파일에 부여된 0과 1 사이의 값을 갖는 정규화된 주제 카테고리 점수(NQcate_score)를 가중치로 반영한 알고리즘으로 순위를 조정하였다. 즉, 입력된 질의어가 이용자의 프로파일과 전혀 관계가 없을 경우에는 가중치가 0이 되어 일반 검색 결과를 제시한다. 반대로 질의어가 이용자 프로파일과 완전 일치할 경우에는 개인화된 검색 결과만을 제시한다.

이 실험에서는 질의어를 랜덤 질의어와 주제 질의어로 나누어 검색한 결과를

개인화 검색에 의해 재순위화하여 검색 성능을 비교하였다. 랜덤 질의어는 중위빈도 질의어 중에서 무작위로 추출하였다.

랜덤 질의어에 대한 검색 실험결과 개인화된 검색에 대한 적합성이 일반 검색 결과에 비해 5.9% 향상된 것으로 나타났다. 개인화 검색 결과가 오히려 좋지 않다고 응답한 이용자는 활동 중인 카페가 없고, 신지식 활동도 아예 없거나 미비하였다. 실험결과에서 보듯이 개인화 검색의 만족도는 이용자의 관심 주제를 얼마나 정확하게 설정할 수 있느냐에 달려있다. 잘못 설정된 관심 주제에 대해서는 이용자가 직접 수정할 수 있는 방법이 반영되어야 할 것이다.

반면 이용자의 관심 주제 카테고리화 관련된 질의어로 검색한 결과 개인화 검색이 일반 검색에 비해 26%의 적합성 향상률을 기록했다. 이 실험을 통해 개인화 검색에서는 자신의 관심 주제와 관련된 질의로 검색했을 때 검색 결과의 적합성이 크게 향상되는 것으로 나타났다.

일반 검색에 비해 개인화된 검색결과의 적합성이 높게 나타난 질의어는 중의적인 단어인 경우와 폭넓은 개념을 갖는 단어였다. 실험용 질의어 중 랜덤 질의어 10개에 대해서는 이 질의어들이 관심 주제와 전혀 관련이 없었기 때문에 성능의 향상이 전혀 없었던 반면 주제 관련 질의어에 대해서는 1대개의 질의어에서 33%-67% 사이의 높은 향상률을 가져온 것으로 나타났다.

4.4 사용자 그룹핑에 의한 추천 주제 선정 실험 결과

주제 카테고리를 확장하여 추천해 주기 위한 기본 조건으로 최소 4개 이상의 카테고리를 가져야 하기 때문에 대상 이용자가 제한되었다. 해당하는 이용자 그룹의 주제 카테고리 가운데 이미 주제 카테고리로 파악된 제외한 나머지 카테고리를 빈도순으로 정렬한 다음 동시 출현 빈도가 평균 이상인 카테고리만을 추천하였다. 추천한 카테고리에 대해서는 이용자가 직접 적합성 여부를 판정하였다.

4.5 추천서비스에 대한 만족도 평가

이용자별 주제 카테고리의 최종 매핑결과를 이용하여 추천한 질의어와 카페에 대해

이용자가 얼마나 만족하고 있는지를 5점 척도로 조사했다. 질의어는 Ucate_score가 가장 높은 카테고리만을 대상으로 하여, NQcate_score가 높은 질의어 10개를 추천하였고, 카페는 NCcate_score가 높은 카페 5개를 추천하였다.

추천 질의어와 추천 카페에 대한 이용자별 만족도를 평가한 결과 질의어나 카페 추천서비스에 대한 만족도는 각각 3.45와 3.65로 나타났다. 이는 5점 척도에서 보통 이상의 만족도라고 볼 수 있으나 개선의 여지가 있다고 판단되는 결과이다. 본 실험에서는 질의어와 카페만을 추천하였지만, 다양한 부가 정보를 추천할 수 있다면 추천 서비스에 대한 이용자의 전반적인 만족도는 향상될 수 있을 것으로 보인다.

5. 결론

한국의 포털 검색 서비스는 전형적인 웹사이트뿐만 아니라 뉴스를 포함한 각종 콘텐츠, 카페 글이나 지식 검색 등 이용자들이 생산하는 콘텐츠 등을 포함하는 매우 포괄적인 서비스의 성격을 띤다. 지금까지 검색 결과의 만족도를 높이기 위한 개인화 서비스가 다각도로 연구되었으나, 대형 포털 환경에 적용 가능하면서 실제로 이용자에게 유용성을 제공하는 개인화 시도는 아직 드물다. 따라서, 본 연구에서는 한국형 커뮤니티 기반 포털에 적합한 확장형 개인화 검색 서비스 모형을 제시하고, 실험을 통해 모형의 유용성을 평가하였다.

먼저 일정기간 입력된 질의어에 대한 주제 카테고리 자동 매핑 결과는 다음과 같다. 다음 신지식 서비스(kin.daum.net)를 이용한 주제 카테고리 매핑은 전문가 그룹에 의한 적합성 판단에서 5점 척도 중 평균 4.25점의 상당히 높은 적합성을 보여주었다. 포털의 특성상 고빈도로 발생하지만 정보나 지식 관련 정보요구가 아닌 일시적인 이슈나 경향과 관련된 질의어와 사이트 찾기성 질의어들이 상대적으로 부적합한 결과를 보였다. 상위빈도 질의어와 하위빈도 질의어간 만족도의 차이는 크지 않았으며, 검색결과 건수가 많을수록 적합성이 높은 것으로 나타났다.

이용자의 카페활동에 따른 개인의 관심 주제 카테고리 선정에서 카페 활동도가 높을수록 개인화된 검색결과 및 추천서비스의 만족도가 높게 나타났다. 이는 커뮤니티 서비스가 즉시적이고 다양한 관심사를 반영하는 검색 서비스보다 더 집중적이고 상대적으로 장기적인 관심 주제를 반영하기 때문인 것으로 보인다.

이용자별로 최종 선정된 주제 카테고리를 기반으로 한 개인화 검색 결과 및 추천 서비스에 대한 평가는 다음과 같다.

첫째, 랜덤 질의어를 대상으로 한 개인화 검색에서는 만족도가 5.9% 향상된 반면, 주제와 관련된 질의어를 대상으로 한 결과에서는 26% 향상되었다. 특별히 개인화 검색의 효용성은 개인의 관심사와 관련된 질의어의 경우에 의미가 있음을 알 수 있었다.

둘째, 이용자 그룹에 의한 주제 카테고리 추천 결과 추천의 정확성은 추천방식에 따라 80%~100%로 나타났다.

셋째, 추천된 질의어에 대한 만족도는 5점 척도에서 평균 3.45, 추천된 카페에 대한 적합성은 평균 3.65로 나타났다. 추천되는 질의어 및 카페가 이용자의 관심사에 적합하더라도 질의어나 카페 추천이 이용자에게 반드시 필요한 서비스는 아닐 수 있다는 것을 시사한다.

이상의 실험 결과를 통해 대용량의 실제 로그 데이터를 토대로 이용자의 관심 주제를 자동으로 추출하는 개인화 검색 서비스 모형이 실제 웹 환경에서 효과적으로 적용될 수 있음을 확인하였다. 또한 주로 검색 이용행태의 분석을 통한 이용자 관심 주제 추출에 국한된 기존 연구와 달리 한국적 환경에 적합한 커뮤니티 서비스를 통한 관심 주제 추출이 효과적임을 알 수 있었으며, 개인화된 검색 결과는 검색 성능의 향상을 가져오는 것으로 나타났다.

참 고 문 헌

김영환. 2002. 『인터넷 사이트 개인화 서비스의 현황과 이용자 반응 조사: 마이 페이지와 뉴스레터 서비스를 중심으로』. 석사학위논문, 서강대학교 언론대학원.

황용석, 이홍철. 2000. 『인터넷 뉴스사이트 편집과 기사 쓰기』. 서울: 한국언론재단.

Bonett, Monica. 2001. "Personalization of Web Services: Opportunities and Challenges." *Ariadne*, 28.

Glover, E. J., S. Lawrence, M. D. Gordon, W. P. Birmingham, and C. L. Giles.

2000. "Web Search - Your Way." *Communications of the ACM*, 44(12): 97-102.
- Jeh, G. and J. Widom. 2003. "Scaling Personalized Web Search." *In Proceedings of the 12th International World Wide Web Conference*, 271-279.
- Johansson, Pontus. 2004. *Design and Development of Recommender Dialogue Systems*. Ph.D. diss., Institute of Technology in Linköping University.
- Liu, Fang, Yu, Clement and Meng. Weiyi. 2002. "Personalized Web Search by Mapping User Queries to Categories." *In Proceedings of the 11th international conference on Information and knowledge management SESSION: Web search 2*, 558-565.
- Perugini, Saverio. 2004. *Program Transformations for Information Personalization*. Ph.D. diss., Virginia Polytechnic Institute and State University
- Riecken, D. (ed.). 2000. "Personalized Views of Personalization." *Communications of the ACM*, 43(8): 27-28.
- Scime, A. and L. Kerschberg. 2000. "WebSifter: An Ontology-Based Personalizable Search Agent for the Web." *In Proceedings of International Conference on Digital Libraries: Research and Practice*, 493-446.
- Shahabi, Cyrus and Yi-Shin Chen. 2003. "Web Information Personalization: Challenges and Approaches." *In Proceedings of 3rd Workshop on Databases in Networked Information Systems (DNIS)*.
- Tanudjaja, F. and L. Mui. 2002. "Persona: A contextualized and personalized web search." *In Proceedings of the 35th Annual Hawaii International Conference on System Sciences*.
- Zhu, S., X. Deng, K. Chen, and W. Zheng. 2001. "Using Online Relevance Feedback to Build Effective Personalized Meta-search Engine." *In Proceedings of Second International Conference on Web Information*

Systems Engineering.

KCS I