질의어 패턴 자동분석을 통한 커뮤니티 기반 개인화 검색

(Personalized Search based on Community through Automatic Analysis of Query Patterns)

박 건 우 † 이 상 훈 ††
(Gun-Woo Park) (Sang-Hoon Lee)

요 약 기존의 웹 검색 엔진들은 사용자의 검색 의도를 충분히 반영하지 못하기 때문에 사용자가 원하는 정확한 정보를 찾기가 어렵다. 따라서 최근에는 개인의 검색 패턴을 분석하여 검색에 반영함으로써 검색 결과에 대한 만족도를 높이기 위한 많은 연구들이 진행되고 있다. 이러한 개인화 검색을 통해 사용자는 방대한 웹상의 정보들 중 자신의 검색 의도에 보다 적합하고 정확한 정보를 획득할 수 있다. 본 논문에서는 웹 사용자들의 질의어 사용 빈도수(Frequency)에 대한 랭킹 정보를 통해 최근 주요 관심사(Interest)를 파악하고, 주요 관심사 별로 형성된 커뮤니티(Community)를 기반으로 수행되는 개인화 검색 방안을 제안한다. 실험결과 질의어 빈도수, 관심사 및 커뮤니티를 검색에 반영할 경우 개인의 검색 의도에 보다 적합한 검색 결과가 제공되는 것을 확인할 수 있다.

키워드 : 개인화 검색, 검색 의도, 관심사, 질의어 빈도수, 커뮤니티

Abstract Since the existing Web search engines don't sufficiently reflect user's search intent, it is very difficult to find out accurate information that users want to find. Therefore, a lot of researches, study for personalized search, to enhance satisfaction of Web search results by analyzing search pattern and applying it to search are in progress in these days. Web searchers can more efficiently find information and easily obtain appropriate information through the personalized search. In this paper, we propose the personalized search based on community through the analysis of web users' query patterns and interest. Consequently, when applying query frequency, interest and community to web search, we are able to the confirm that the search results which hit to the search intent of the individual are provided.

Key words : Personalized Search, Search Intent, Interest, Query Frequency, Community

1. 서 론

기존의 웹 검색 엔진들은 일반 대중에 의해 많이 참조되는 웹 문서에 동일한 랭킹(Ranking)을 부여하여 웹 사용자들에게 일관적으로 제공한다. 랭킹을 부여하기 위해 문서 링크 수, 사용자 방문내역(click-through history) 등에 대한 정보를 사용하며 구글 검색엔진에서 사용하고 있는 대표적인 알고리즘으로 페이지 랭크(Page-Rank) 알고리즘이 있다. 하지만 TREC(http://www.nist.gov)의 연구 결과를 통해 페이지 랭크 알고리즘을 검색에 적용하는 것은 경우에 따라서 보다 좋지 못한 성능을 나타내는 것을 알 수 있다[1-4]. 이는 실제 검색에 사용되는 질의어에 대해 웹 사용자들의 검색 의도(Search Intent)를 무시한 채 내용 기반의 문서검색에 대한 질의 형태만을 고려하였기 때문이다. 즉 기존의 웹 검색은 사용자의 질의 의도를 충분히 반영하지 못한다는 단점으로 인해 사용자의 주요 관심사(Interest)에 적합한 검색 결과를 획득하기에 제한사항이 있다.

· 이 논문은 2008 한국컴퓨터종합학술대회에서 '질의어 패턴 자동분석을 통한 커뮤니티 기반 개인화 검색'의 제목으로 발표된 논문을 확장한 것임

† 정 회 원 : 국방대학교 국방관리대학원 전산정보학과
pgw4050@hotmail.com
†† 종신회원 : 국방대학교 국방관리대학원 전산정보학과 교수
pgw4050@hotmail.com

논문접수 : 2008년 8월 27일
심사완료 : 2009년 5월 22일

Copyright©2009 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지: 데이타베이스 제36권 제4호(2009.8)

웹 사용자가 만족할 수 있는 개인화된 검색을 실현하기 위해서는 사용자가 입력하는 질의어에 대한 정확한 의미를 파악하고 사용자의 성향 및 주요 관심사에 따라 필요로 하는 정보가 무엇인지 알아야 한다. 따라서 개인화에 부합된 웹 검색을 위해서는 질의어의 특성을 분석하고 검색 의도와 개인의 주요 관심사를 파악할 필요성이 있다. 이를 위해 웹 사용자들의 질의어 패턴 분석 즉, 질의어 랭킹 정보를 통해 웹 사용자들의 최근 주요 관심사를 파악 할 필요성이 있으며, 주요 관심사 별 커뮤니티(Community)를 형성하여 검색에 반영함으로써 사용자의 검색 의도와 관심사에 보다 더 근접한 검색 결과를 획득할 수 있을 것이다. 커뮤니티(Community)의 정의는 "평소 관심사가 비슷하거나 같은 네티즌들이 직접 정보를 생산, 공유하고 이들이 모여 활동할 수 있는 인터넷상의 공간"이다.

본 논문에서는 웹 사용자들의 질의어 패턴 분석하고 주요 관심사를 기반으로 형성된 커뮤니티를 통해 검색을 수행하는 방식의 개인화 검색 시스템을 제안한다. 제안한 검색 시스템은 웹 사용자의 질의어에 대한 사용 빈도수를 자동으로 확인 및 랭킹을 부여 하여 데이터베이스로 구축한 후 질의어에 대한 랭킹 정보를 통해 개인의 관심사를 파악한다. 질의어에 대한 빈도수 정보는 웹 사용자의 관심사를 파악하는 데에 중요한 정보로 사용되며 상위에 랭크 된 질의어들에 대한 정보는 커뮤니티를 형성하기 위한 기준이 된다. 형성된 커뮤니티들은 검색 과정에서 주요 관심사에 따른 검색 의도를 보다 효과적으로 반영함으로써 개인화 검색이 가능하다.

2. 관련연구

2.1 적절성 피드백(Relevance Feedback)

최근 개인화된 웹 정보검색에 대한 많은 연구가 진행되어 왔다[5-8]. 그 예로 구글의 개인화 검색(Google Custom Search, Google personal, http://labs.google.com/personalized)은 웹 사용자의 관심사항에 대한 분류 목록 선택을 통해 웹 사용자의 개별 관심사항을 입력하도록 요구하고, 이 목록을 검색 결과와 대조하여 개인화된 서비스를 제공한다. 또한 개인 프로파일(Personal Profiles)은 개인화된 페이지랭크 기법에서 질의어와 무관하게 관련 웹 페이지의 가중치를 도출하기 위해 사용되기도 한다[9]. 이와 같이 웹 사용자의 의도에 대한 정보는 검색결과의 적절성에 대한 웹 사용자의 판단을 피드백하는 방법을 통해 수집될 수 있다.

2.2 웹 검색에서의 질의 유형

웹 사용자 질의는 의도에 따라서 세 가지로 구분된다.

- 내용 검색(Informational Need)
- 사이트 검색(Navigational Need)
- 서비스 검색(Transactional Need)

내용 검색은 웹 사용자가 알고자 하는 정보와 관련 있는 순서대로 순위화 된다. 예를 들어 "What is a prime factor?" 혹은 "prime factor"와 같은 질의는 웹 사용자가 "prime factor"에 대해 알기를 원할 것이다. 반면 사이트 검색은 웹 사용자가 찾고자 하는 사이트의 순서대로 순위화 된다. 예를 들어 "Where is the site of John Hopkins Medical Institutions?"나 "John Hopkins Medical Institutions"와 같은 질의를 통해 해당 사이트에 방문하기를 원할 경우 비록 그 사이트와 관련된 문서일지라도 사이트의 중심 출입 문서만을 올바른 문서로 채택한다. 서비스 검색은 웹 사용자가 찾고자 하는 서비스를 제공하는 순서대로 순위화 된다. 예를 들어 "Where can I buy concert tickets?"나 "buy concert tickets"와 같은 질의는 웹 사용자가 콘서트 티켓을 구매할 수 있는 웹 문서를 찾기를 바란다. 이때 검색 엔진은 단순히 질의어와 의미상으로만 관련된 문서가 아닌 웹 사용자가 원하는 서비스 받을 수 있는 가장 적절한 웹 문서를 순위화하여 제공해 준다. 이와 같이 검색 대상과 웹 사용자의 의도가 다양해짐에 따라 각 상황에 적합한 문서를 순위화시키는 정책이 달라져야 한다.

3. 질의어 패턴 분석 및 개인화 검색

제안하는 시스템은 개인화 검색을 위해 웹 사용자들이 사용하는 질의어를 데이터베이스에 저장한 후 각 질의어 사용 빈도수에 따라 랭킹을 부여한다. 시스템은 질의어 랭킹 정보(Query Ranking Information)를 참조하여 웹 사용자의 주요 관심사를 파악한다. 즉 웹 사용자에 의해 일정 기간 반복적으로 사용된 질의어 패턴과 데이터베이스에 축적된 질의어의 랭킹 정보는 웹 사용자의 주요 관심사를 파악하는데 중요한 정보로 사용된다. 또한 특정 임계치(Threshold) 이상의 질의어에 대한 랭킹 정보를 기반으로 커뮤니티를 형성하고, 커뮤니티를 검색에 반영함으로써 웹 사용자의 관심분야와 검색 의도에 좀 더 부합된 개인화 검색 결과를 획득할 수 있도록 한다. 그림 1에 나타나듯이, 개인화 검색을 위해 질의어 빈도수, 검색 의도, 사용자 관심사의 세 가지 요소가 고려되며, 각 요소 간의 상호 관련성에 대한 가중치를 부여하여 검색을 수행함으로써 개인의 검색 만족도를 향상시킬 수 있다.

그림 2는 웹 사용자의 질의어 패턴을 자동으로 분석한 후 질의어에 대한 랭킹 정보를 통해 주요 관심사 별 커뮤니티를 형성하고 이를 기반으로 개인화 검색을 수행하기 위한 전반적인 과정을 나타낸다.

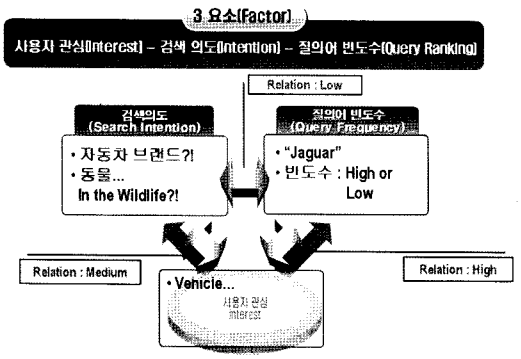


그림 1 개인화 검색을 위한 세 가지 주요 고려 요소

개인화 검색을 위한 각 과정 별 세부적인 처리 내용은 다음과 같다.

3.1 프로그램 설치 운용

각 웹 사용자들의 PC에는 질의어 패턴을 자동으로 분석하고 검색 옵션에 따른 추천 질의어를 조합하여 검색엔진과 연동 운용하기 위해 자체 개발한 프로그램인 서치 브리지(Search Bridge) 프로그램을 설치 운용한다.

그림 3은 웹 사용자들의 PC에 설치 운용되는 서치 브리지 프로그램의 실행 화면이다. 웹 사용자들이 입력한 질의어를 저장하기위한 데이터베이스는 MDB를 설치 운용하며, 각 PC에 저장된 질의어에 대한 모든 정보는 서치 브리지 프로그램에 의해 수집되어 통합 데이터베이스 서버(My SQL Server)에 저장된다.

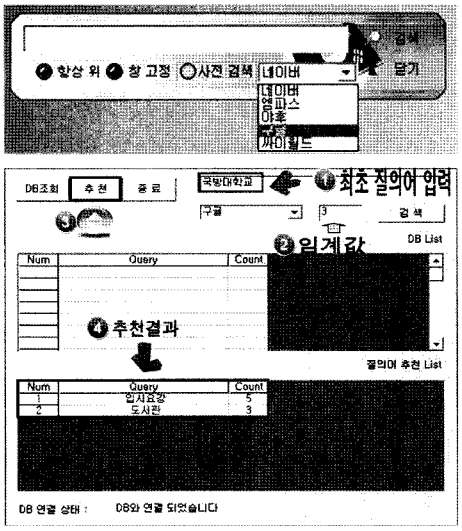


그림 3 서치 브리지 실행 화면

3.2 질의어 추출 및 데이터베이스 생성

웹 사용자가 정보를 검색하기 위해 질의어를 서치 브리지에 입력하게 되면 서치 브리지는 사용자 PC의 MDB 내에 새로 입력된 질의어의 존재여부를 확인한다. 이때 새로 입력한 질의어가 이미 데이터베이스에 존재하면 질의어 사용 횟수에 대한 정보만을 갱신하고 다음 과정을 수행하며, 존재하지 않으면 MDB에 새로운 질의어를 추가한다.

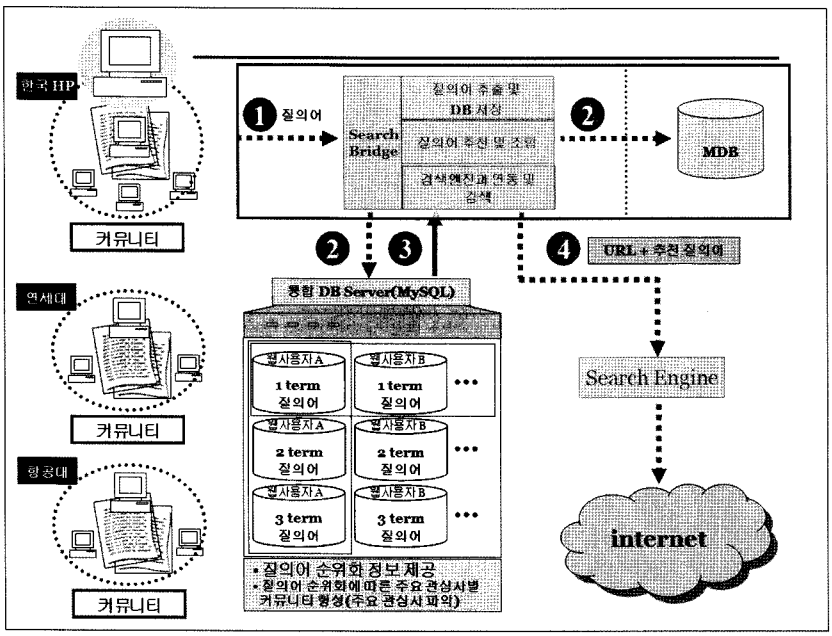


그림 2 질의어 패턴 분석 및 커뮤니티 기반 개인화 검색 구성도

각 웹 사용자 별 저장되어있는 질의어 정보는 서치 브리지에 의해 주기적으로 통합 데이터베이스 서버에 전송되며 이와 같은 과정을 통해 네트워크 내 존재하는 개인 및 전체 웹 사용자들의 질의어에 대한 데이터베이스가 구축된다.

웹 사용자들이 입력하는 질의어는 카운트되며 질의어 빈도수를 기준으로 질의어에 대해 랭킹을 부여하는 과정이 수행된다. 즉, 질의어에 대해 데이터베이스가 구축되면 각 사용자 별, 전체 사용자 별 질의어 빈도수에 대한 랭킹이 부여된다. 질의어 수(Query Size)는 최대 3을 초과하게 되면 오히려 검색 효율이 감소한다는 연구 결과에 따라 질의어 수는 3개까지 고려하여 데이터베이스 구축 후 검색에 활용한다. 그림 4는 질의어 개수 별 데이터베이스 구축 예를 나타낸다.

- 1개의 질의어에 대한 데이터베이스 생성 저장
- 2개의 질의어에 대한 데이터베이스 생성 저장
- 3개의 질의어에 대한 데이터베이스 생성 저장

입력하는 질의어에 대해 공백을 근거로 질의어가 구분되지만 데이터베이스에 저장될 때는 하나의 질의어로 간주하여 데이터베이스를 구축한다. 예를 들어 "정보검색 검색"과 "정보검색 개인화검색"은 서로 다른 질의어로써 검색 시에는 각각 2개의 질의어로 인식되어 검색에 적용 된다. 하지만 데이터베이스에 저장될 때는 각각 하나의 질의어로 인식하여 저장되며 사용자가 동일한 질의어 "정보검색 검색"을 입력하게 되면 기존 데이터베이스를 확인하여 빈도수만 증가시킬 것인지 데이터베이스에 새로운 질의어로 추가 저장할 것인지를 판단한다. 이와 같은 방법으로 최대 3개의 질의어에 대해 각각 다른 데이터베이스를 구축하여 질의어의 빈도수에 따라 랭킹이 부여된다.

3.3 질의어 추천 및 조합

질의어 추출 및 데이터베이스 생성과정에서 파악된 질의어 랭킹 정보를 통해 웹 사용자들의 질의어 선호도 및 주요 관심사를 파악할 수 있다. 이때 웹 사용자들의 주요 관심사 별 커뮤니티를 형성할 수 있으며, 형성된 커뮤니티는 정보 검색에 중요한 요소로 활용된다. 즉 상위 랭크된 질의어는 서치 브리지를 통해 웹 사용자가 입력 한 질의어와 함께 조합되어 검색 엔진에 전송된다. 또한 형성된 커뮤니티를 통해 자신의 관심사항과 유사한 커뮤니티 내에서 자주 사용되는 질의어 즉 상위에 랭크 된 질의어를 추천받아 검색에 이용된다.

3.4 검색

웹 사용자가 최종 검색하는 단계로 선택 검색을 한다. 검색에는 일반적인 검색을 포함하여 다음과 같이 3가지 선택 옵션이 있다.

- 일반 검색: 단순 질의어와 URL의 조합("질의어 + URL"을 서치 브리지에 전송하여 검색)
- 개인화 검색: 개인 PC의 데이터베이스에 저장되어 있는 질의어 순위화 정보를 참조하여 추천된 질의어와 URL 조합("개인의 상위 순위화 질의어 + URL"을 서치 브리지에 전송하여 검색)
- 커뮤니티 검색: 통합 데이터베이스 서버 내 상위 순위화된 질의어 정보를 참조로 형성된 커뮤니티를 기반으로 추천된 질의어와 URL 조합("커뮤니티의 상위 순위화된 질의어 + URL"을 서치 브리지에 전송하여 검색)

4. 실험 및 평가

4.1 데이터 셋

웹 사용자들은 다양한 프로파일(예: 나이, 사는 지역, 결혼 여부, 직종, 관심사, 학교, 전공 등)을 갖고 있는 500명을 대상으로 자체 개발한 서치 브리지를 설치 운용 하도록 하여 일정 기간(2008. 3~2008. 6) 질의어 사용에 대한 정보를 수집하였다.

4.2 질의 빈도수와 관심사의 관계 분석

그림 5는 각각 관심분야(예: 전자회로, 노트북)가 서

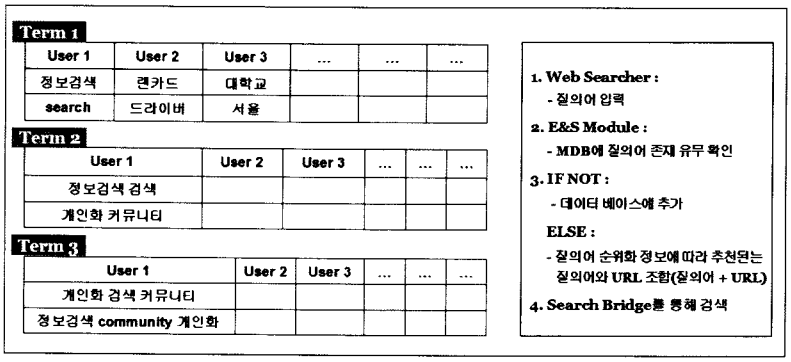
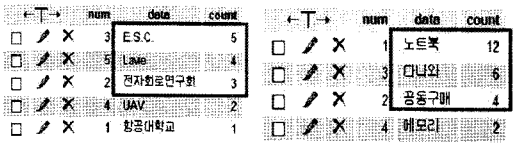


그림 4 질의어 수별 데이터베이스 구성

로 다른 웹 사용자들의 질의어에 대한 순위화 정보를 나타낸다. 그림에 나타나듯이 상위에 순위화된 질의어들은 평소 관심분야와 매우 연관성이 높은 것을 확인할 수 있다. 즉 웹 사용자들이 빈번하게 사용하는 질의어는 그 사람의 평소 관심사항을 잘 대표한다고 할 수 있다. 따라서 상위 순위화 되어있는 질의어를 통해 웹 사용자의 질의어 패턴을 분석하고 이를 기반으로 커뮤니티를 형성하여 정보 검색에 활용하면 보다 개인의 관심사에 부합하고 의도하는 정보를 효과적으로 획득할 수 있다.

그림 6은 모두 평소 야생 동물에 관심이 많은 웹 사용자가 질의어 "jaguar"를 입력하여 검색한 결과화면이다. 그림 6의 (a)는 구글 검색엔진을 이용하여 일반 검색을 한 결과로써 웹 사용자의 평소 관심사와 검색 의도와는 무관한 자동차 브랜드와 관련된 정보들이 웹 문서의 상위에 순위화되어 제공되어지는 것을 확인할 수 있다. 하지만 그림 6의 (b)는 구글 검색엔진으로 검색하기 전에 서치 브리지를 통해 필터링을 거쳐 검색한 결과이다. 그림 6과 달리 야생 동물에 관련된 정보가 상위에 순위화되어 제공됨으로써 평소 웹 사용자의 관심사와 검색 의도가 반영되는 것을 확인할 수 있다.

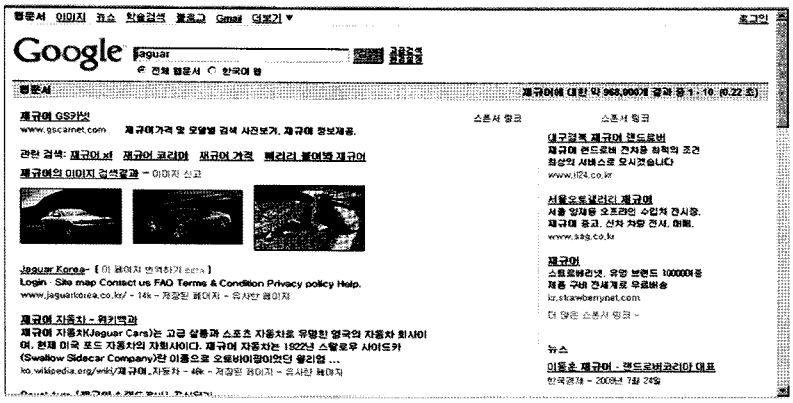


(a) 전자공학과 학생 (b) 노트북에 관심 있는 학생

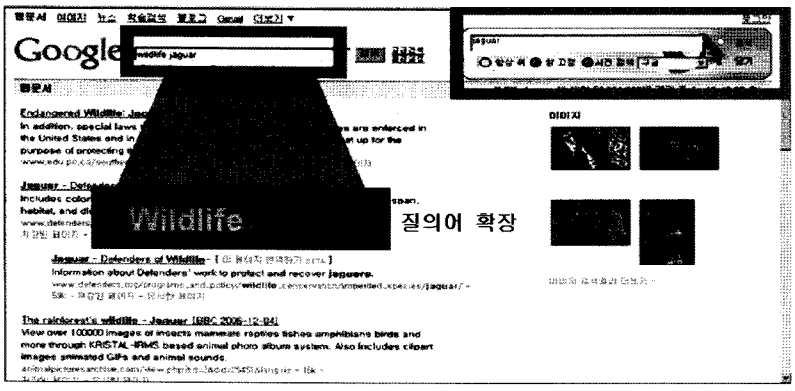
그림 5 사용 질의어에 대한 빈도수 별 순위화

5. 결론 및 향후 연구

개인의 관심사에 적합한 정보를 보다 효율적으로 검색하기 위해서는 사용자의 검색 패턴과 주요 관심사를 파악하는 것이 중요하다. 따라서 본 논문에서는 웹 사용자들의 질의어 패턴을 자동으로 분석하고 분석된 질의어 패턴에 대한 순위 정보를 기준으로 주요 관심사항을 파악하여 검색에 활용하였다. 또한 상위에 순위화된 질의어 정보를 통해 주요 관심사 별 커뮤니티 형성이 가능



(a) 구글을 이용한 검색



(b) 서치 브리지를 이용한 검색

그림 6 질의어 "jaguar"를 입력하여 검색한 결과화면

하며 이와 같은 방법으로 형성된 커뮤니티를 검색에 반영함으로써 검색을 보다 용이하고 효과적으로 할 수 있음을 알 수 있다. 결론적으로 개인의 정보검색 패턴을 검색에 반영하고 자신과 관심분야가 유사한 커뮤니티를 검색에 반영함으로써 보다 효과적인 검색 결과를 획득할 수 있었다.

향후에는 더 많은 다양한 분야의 웹 사용자들에게 서치 브리지를 적용하여 질의어 패턴을 분석할 필요성이 있다. 또한 현재까지는 질의어에 대한 상위 순위화된 정보를 휴리스틱한 방법으로 주요 관심사 별 커뮤니티를 형성하여 정보검색에 이용하였으나 향후에는 자동으로 커뮤니티를 형성할 수 있는 방안에 대한 연구를 수행할 것이다. 뿐만 아니라 커뮤니티 간 소셜 네트워크(Social Network)를 형성하여 타 관심분야의 웹 사용자들에 대한 검색 패턴을 상호 연계하여 정보검색에 활용할 수 있는 방안을 연구할 것이다.

참 고 문 헌

[1] Croft, W.B., "Combining Approaches to Information Retreval : Recent Research from the Center for Intelligent Information Retrieval," *Kluwer Academin Publishers*, pp.1-36, 2000.

[2] Brin, S. and Page, L., "The Anatomy of a Large-scale Hypertextual Web Search Egine," *Computer Nettworks and ISDN Systems*, vol.30, no.1-7, pp.107-117, 1998.

[3] Craswell, N., Hawking, D. and Robertson, S., "Effective Site Finding using Link Anchor Information," *In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, LA, pp.250-257, 2001.

[4] Yang, K., "Combining Text and Link-Based Retrieval Methods for Web IR," *In Text Retrieval Conference(TREC-10)*, Gaithersburg, Maryland, pp.609-618, 2001.

[5] S. Lee, Y. Chung, "Design and Evaluation of a Personalized Search Service Model Based on Web Portal User Activities," *Journal of KOSIM*, vol.23, no.4, pp.179-196, Dec. 2006. (in Korean)

[6] J. Lee, S. Cheon, "A Personalized Search based on Query Expansion," *Proc. of the KIISE 2008*, vol.35, no.2(C), pp.203-208, Oct, 2008. (in Korean)

[7] Barry Smyth, "A Community-Based Approach to Personalizing Web Search," *IEEE COMPUTER SOCIETY*, vol.40, Issue.8, pp.42-50, Aug. 2007.

[8] Alessandro Micarelli, Fabio Gasparetti, Filippo Sciarrone, Susan Gauch, "Personalized Search on the World Wide Web," *LNCS*, pp. 195-230, vol.4321, 2007.

[9] Cantador, I., Castells, P. "Extracting Multilayered Semantic Communities of Interest from Ontology-based User Profiles: Application to Group Modelling and Hybrid Recommendations. Computers in Human Behavior," *Elsevier*, Special issue on Advances of Knowledge Management and the Semantic Web for Social Networks, 2008.

박 건 우

1997년 충남대학교 컴퓨터과학과 졸업(학사). 2007년 연세대학교 컴퓨터과학과 정보통신 연구실 졸업(석사). 2007년~현재 국방대학교 전산정보학과 재학(박사과정). 관심분야는 정보검색, 소셜 네트워크, 네트워크, 네트워크 보안

이 상 훈

1978년 성균관대학교 정보통신공학과 졸업(학사). 1989년 연세대학교 산업대학원 전산학과 졸업(석사). 1997년 일본 교토대학교 정보공학 졸업(박사). 1998년~2000년 서일대학 겸임교수, 충남산업대학교 교수. 2000년~현재 국방대학교 전산정보학과 교수. 관심분야는 정보검색, 멀티미디어 데이터베이스, HCI