https://doi.org/10.5392/JKCA.2017.17.02.113

소셜 미디어 사용자의 최근 관심사를 고려한 소셜 검색 기법

Social Search Scheme Considering Recent Preferences of Social Media Users

송진우, 전현욱, 김민수, 김기훈, 노연우, 임종태, 복경수, 유재수
충북대학교 정보통신공학과

JinWoo Song(jinwoo825@chungbuk.ac.kr), Hyeonwook Jeon(wooky@chungbuk.ac.kr),
Minsoo Kim(mskim88@chungbuk.ac.kr), Gihoon Kim(gihoon kim@chungbuk.ac.kr),
Yeonwoo Noh(ywnoh@chungbuk.ac.kr), Jongtae Lim(jtlim@chungbuk.ac.kr),
Kyoungsoo Bok(ksbok@chungbuk.ac.kr), Jaesoo Yoo(yjs@chungbuk.ac.kr)

요약

기존의 소셜 검색은 사용자의 프로파일의 최신성과 유사한 사용자의 관심사를 고려하지 않기 때문에 검색 결과가 사용자에게 적합하지 않다는 문제가 있다. 이에 따라 시간적 속성과 다른 사용자의 관심사를 고려한 소셜 검색 연구가 요구되고 있다. 본 논문에서는 시간에 따른 최근 관심사, 사용자와 유사도가 높은 사용자들의 관심사를 고려한 소셜 검색 기법을 제안한다. 제안하는 기법은 사용자의 최근 관심사를 고려하기 위해 소셜 미디어 사용자의 활동 정보를 분석한다. 다른 사용자들의 관심사를 분석한 정보와 결합하여 랭킹을 수행함으로써 검색 결과의 만족도와 정확성을 향상시킨다. 성능평가를 통해 제안하는 소셜 검색 기법이 기존 기법에 비해 성능이 우수함을 보인다.

■ 중심어 : | 소셜 검색 | 개인화 검색 | 소셜 미디어 | 소셜 네크워크 | 관심사 |

Abstract

The existing social search has a problem that search results are not suitable for a user since it does not take into account the recency of the user profile and the interests of similar users. Therefore, studies on a social search considering a temporal attribute and the interests of other users are required. In this paper, we propose a social search scheme that takes into account the recent interests of a user by time and the interests of the most similar users. The proposed scheme analyzes the activity information of a social media user in order to take into account the recent interests of the user. And then the proposed scheme improves the satisfaction and accuracy of search results by combining the interests of similar users with the analyzed information and performing ranking, It is shown through performance evaluation that the proposed scheme outperforms the existing scheme.

■ keyword : | Social Search | Personalized Search | Social Media | Social Network | Preference |

* 이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원(No. 2016R1A2B3007527)과 미래창조과학부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업의 연구결과로 수행되었음(IITP-2016-H8501-16-1013)
접수일자 : 2016년 10월 28일
수정일자 : 2016년 12월 05일
심사완료일 : 2016년 12월 05일
교신저자 : 유재수, e-mail : yjs@chungbuk.ac.kr

I. 서 론

태블릿PC, 스마트폰 등과 같은 모바일 기기의 등장과 함께 무선 네트워크 환경이 발달하면서 이로 인해 온라인으로 사용자들의 인맥을 관리하고 정보를 상호 교환하는 소셜 미디어(social media)들이 등장하게 되었다[1][2]. 일반적인 전통 매체는 일대다 관계로 정보 전달이 목적이었지만, 소셜 미디어는 다대다 관계로 쌍방향 관계를 가지며 상호간의 다양한 정보 공유와 관계를 형성한다. 사용자들은 소셜 활동을 이용하여 사용자들의 현재 관심분야가 무엇인지, 어떤 사람과 관계를 맺고 있는지를 분석할 수 있다. 소셜 검색에서 게시물 생성, 북마크, 코멘트, 점수평가 등의 피드백을 활용하면 그 사용자들의 관심사에 대한 프로파일(Profile)을 얻어낼 수 있다[3][4]. 최근 많은 사람들의 소셜 미디어 활동으로 인해 소셜 데이터가 급증하고 있으며, 이는 사용자의 성향을 판단하는데 유용하게 사용된다. 대량의 소셜 데이터들은 사용자의 관심사를 판단하여 개인화된 소셜 검색을 제공하기에 적합하다.

일반적인 웹 사용자는 자신이 원하는 결과들을 리스트 최상위에 보여주기를 희망한다. 그러나 웹 검색 결과의 대부분은 사용자 개개인이 원하는 맞춤형 정보가 제공되지 않는다. 이는 실제 검색에 사용되는 질의어에 대해 웹 사용자들의 검색 의도를 무시한 채 내용 기반의 문서검색에 대한 질의 형태만을 고려하기 때문이다. 즉 기존의 웹 검색은 사용자의 질의 의도를 충분히 반하지 못 한다는 단점으로 인해 사용자의 주요 관심사에 적합한 검색 결과를 획득하기에 제한사항이 있다. 예를 들어 웹에서 '리버풀'을 검색하면, '축구팀 리버풀FC'와 '도시 리버풀'에 대한 결과가 혼재되어 있다. 보편적으로 '축구팀 리버풀FC'가 상위랭크에 노출되어 있으며 표출되는 사이트의 수도 상대적으로 많다는 것을 확인할 수 있다. 웹에서 이러한 결과를 보여주는 이유는 정보제공을 하는 웹이 단어의 모호함, 중의성을 고려하여 대중들이 가장 많이 찾는 결과물을 보여줘야 하기 때문이다. 즉, 대부분의 사용자가 '축구팀 리버풀FC'로 표현된 사이트를 자주 클릭하는 경향 을 보이기 때문에 상대적으로 많은 웹페이지를 보여주고 있다.

사용자들은 전통적인 방식의 웹 검색이 아닌 사용자 개인의 성향과 관심사를 고려한 소셜 검색의 요구가 증가하고 있다. 기존의 웹 검색 방식은 사용자가 원하는 결과를 최상위에 보여주지 못한다. 또한 많은 웹 검색 결과 중에서 사용자들은 원하는 정보를 얻거나 원하지 않는 정보를 피하기 위해 더 많은 노력을 해야 한다. 이런 문제를 해결하기 위해서는 사용자가 원하는 정보를 가진 문서를 제공하며 그 결과를 상위에 위치시키는 검색 방법이 필요하다. 사용자가 만족할 수 있는 소셜 검색을 실현하기 위해서는 사용자가 입력하는 질의어에 대한 정확한 의미를 파악하고 사용자의 성향 및 주요 관심사에 따라 필요로 하는 정보가 무엇인지 알아야한다. 따라서 소셜 검색을 위해서는 질의어에 대한 의미를 알아야하고 사용자의 관심사를 파악할 필요성이 있다.

소셜 검색은 사용자의 관심사를 바탕으로 원하는 정보를 위주로 검색 결과의 상위에 노출시키는 목적으로 사용된다. 소셜 검색을 위해서는 앞서 언급한 것처럼 사용자들의 관심사를 파악해야한다. 관심사를 파악하기 위해선 사용자들의 활동 내역을 알아야하는데 이는 소셜 미디어의 활동 내역으로 알 수 있다. 소셜 미디어의 활동 내역을 텍스트 마이닝 기법을 활용하여 사용자가 어떠한 카테고리에 관심이 있는지 판단할 수 있고 이를 이용해 사용자가 관심 있는 검색 결과를 상위에 배치시킬 수 있으며 사용자들의 공통 관심사를 이용한 방법으로 사용자들에게 원하는 검색 결과를 제공할 수 있다.

본 논문에서는 소셜 미디어 환경에서 사용자의 최근 관심사와 유사한 사용자의 관심사를 고려한 개인화 검색 기법을 제안한다. 제안하는 기법은 사용자들의 소셜 미디어 활동을 텍스트 마이닝 기법으로 분석하여 사용자의 관심 카테고리를 구축하며 관심 카테고리를 고려한 검색을 제공한다. 그리고 사용자들의 소셜 관계를 고려하여 사용자들의 프로파일 정보와 유사도가 높은 사용자, 특정 카테고리의 전문가 그리고 친구로 연결된 사용자들의 관심사를 고려한 검색을 제공한다. 또한 사용자의 관심사를 시간에 따른 가중치를 두어 가장 최신

의 관심사를 고려하였다. 이러한 소셜 활동으로 얻은 개인의 관심사를 활용하여 관심 카테고리의 일치 정도에 따른 랭킹 검색 기법을 제안한다. 제안하는 기법은 기존의 소셜 검색 기법과 다르게 사용자의 최근 관심사를 활용하여 관심사의 최신성을 유지하고 유사도가 높은 사용자들의 관심사를 고려하여 검색 결과의 만족도와 정확성을 향상시킨다. 제안하는 기법으로 도출된 검색 결과의 만족도와 정확성의 향상을 확인하기위해 만족도를 평가하는 nDCG(Normalized Discounted Cumulative Gain)와 정확성을 평가하는 정확도(precision)와 재현율(recall) 이용하여 두 가지 방법으로 평가한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 관련된 연구를 소개하고 문제점을 제시한다. Ⅲ장에서는 제안하는 기법의 특징과 과정에 대해 기술한다. Ⅳ장에서는 제안하는 기법의 우수성을 실험을 통해 검증한다. 마지막으로 Ⅴ장에서는 논문의 결론에 대해 기술한다.

Ⅱ. 관련연구

웹 검색의 효율성을 높이는 연구는 기존에도 활발히 연구되어 왔다. 특히 검색 기록 분석을 이용하여 성향 정보를 파악하고 이를 다시 검색 결과에 반영하여 개인의 관심 정보를 효율적으로 제공하는 방안들이 연구되어 왔다. B. Carterette은 기존의 질의와 문서의 연관성에만 의존한 검색 엔진의 평가 방식을 개선하기 위해 사용자의 클릭 로그에 따른 문서의 만족도를 측정하였다. 웹 검색 결과의 클릭 로그를 활용하여 사용자들의 행위를 토대로 사용자의 성향을 모델링하고 nDCG(normalizing Discounted Cumulative Gain)를 통해 사용자의 만족도를 평가하는 논문을 제시했다[5]. 그러나 사용자의 최근 관심사를 고려하기 어렵고 클릭 로그를 사용하여 관심사를 모델링하기 때문에 관심사 모델링이 페이지 결과에 의존하게 되어 관심 카테고리의 특징을 확연하게 표현하는 데에는 어려움이 있다. [6]에서는 기존의 정보를 얻던 방식보다 더 나은 사용자, 문서, 대답을 얻기 위해 질의에 대한 전문가를 찾는 시스템이 제안되었다. 사용자가 질의를 보내면 사용자의 소셜 미디어를 분석해 사용자의 순위를 결정하여 질문들을 전달해주는 소셜 질의/답변 시스템을 제안되었다. 하지만 사용자가 활동한 내용을 모두 종합하여 현재 사용자의 질의에 대한 결과 값을 제공하기 때문에 사용자의 최근 관심사를 파악하기엔 어려움이 있다. [7]에서는 많은 소셜 미디어 서비스가 증가함에 따라 선호하는 웹 페이지를 공유하기위해 소셜 주석을 이용하여 웹 문서의 성향을 파악하고 그 웹페이지의 인기도를 측정하는 기법이 제안되었다. 사용자의 질의와 소셜 주석의 유사도를 표현하여 검색결과를 제공한다. 그러나 사용자의 질의와 소셜 주석의 유사도가 떨어지는 경우에 적절한 검색을 제공할 수 없다. 그리고 일반적인 웹 페이지는 문서의 카테고리를 알 수 없기 때문에 제안한 기법에 적용할 수 없는 문제점을 가지고 있다.

초창기에는 소셜 웹 검색이 클릭 로그를 통해 연구되었지만 최근 논문에는 소셜 미디어를 활용하여 소셜 검색 기법을 연구하고 있다. [8]에서는 유사한 그룹의 사용자들의 관심사와 피드백을 기반으로 하는 웹 검색 결과를 제공하는 소넷랭크(SonetRank)가 제안되었다. 소넷랭크는 사용자의 문서에 대한 선호도, 질의와 관련된 소셜 그룹 사용자 및 네트워크의 다른 사용자들을 활용한다. [9]에서는 사용자의 프로파일을 작성할 때 기록해 두었던 정적인 관심사에 기반 하여 그에 해당하는 카테고리를 구축하고 사용자의 선호도와 관심사 키워드들을 웹 검색 결과와 일치시켜 내림차순으로 정렬한 검색 결과를 제공하는 방법이 제안되었다. 그러나 대부분의 사람들은 관심사가 수시로 변경되며 그에 따른 검색의 성향도 달라진다. 또한 자신과 비슷한 다른 사람들의 관심사를 적용하고 있지 않기 때문에 기존의 검색 기법들로 사용자들에게 한정된 검색내용을 제공할 수 밖에 없다. 이러한 이유 때문에 기존의 검색 기법은 만족성이 저하되고 정보의 최신성이 반영되지 못하는 문제점과 다른 사용자들의 최근 성향을 반영하지 못하는 문제점이 있다.

소셜 미디어에서 자신과 가까운 사람이나 영향력이 큰 사람을 검색함으로써 신뢰할만한 사용자를 검색하는데 유용하다. 또한 공통 관심사를 이용하면 자신과 주변 사용자들의 관심사를 고려하여 개인화된 문서를

검색하는데 유용하다. 앞서 언급한 것처럼 많은 연구들을 통해 정확하고 효율적인 사용자의 관심사 추출이 가능해졌다. 또한 이를 개선하기 위한 연구들도 많이 진행이 되고 있다. 하지만 기존 연구들은 사용자의 행동을 통해서만 관심사를 추출하기 때문에 다른 사용자의 관심사를 표현하기에는 어려움이 있다[10]. 그리고 사용자의 과거부터 현재까지의 모든 관심사를 고려하여 검색 결과를 제공하기 때문에 관심사의 최신성을 반영하지 못한다.

대부분의 사람들은 관심사가 시간이 지남에 따라 변경되며 그에 따른 검색의 성향도 달라진다. 최근에는 소셜 미디어를 활용하여 시간의 흐름에 따라 바뀌는 사용자들의 관심사와 성향을 파악하고 이를 반영한 소셜 검색들이 연구 되고 있다[11-13]. [11]에서는 소셜 북마킹 서비스를 이용해 각 사용자의 최근 관심사를 얻고 답변자의 최근 관심사를 반영해 질문에 가장 잘 대답해 줄 수 있는 잠재 답변자를 검색하는 방식으로 최근 사용자 관심사를 고려한 소셜 검색 알고리즘이 제안되었다. 이 논문에서는 사용자의 최근 관심사를 del.icio.us에서 제공하는 북마킹 태그와 기존 토픽을 병합하여 얻었다. 이는 사용자의 최근 관심사가 어떠한 관심사인지 북마킹 태그와 웹사이트의 특성으로 사용자의 관심사를 간접적으로 알 수밖에 없고 정확한 관심사를 찾아낸다고 보기는 어렵다. 또한 최근 관심사를 5일이라는 기준을 두어 비교적 가장 최근 관심사에만 집중하였고 시간에 따른 가중치를 두지 않아 5일 전의 관심사와 오늘의 관심사에 대해 비중을 두지 않았다.

[12]에서는 사용자들의 최근 관심사를 시간에 따라 분류한 프로파일을 제안하였고 소셜 미디어의 네트워크적 특성을 고려한 개인화 검색이 제안되었다. 시간에 따라 관심 토픽이 변하는 프로파일을 제안하였으나 비교적 과거와 가장 최근 시간대에 따른 시간 가중치가 없으며 정확한 시간 기간을 두지 않았으며 소셜 미디어에서 실시간으로 생성되는 사용자의 성향과 관심사를 반영하지 않는다는 문제점이 있다.

[13]에서는 스카이라인을 활용한 암시적인 정보 수집 기법을 통해 신뢰성을 향상시키고 장소에 대한 성향 정보를 피드백 받아 검색에 반영하는 소셜 검색 기법이 제안되었다. 그러나 위치기반의 소셜 검색이기 때문에 사용자의 관심사 보단 방문정보, 방문위치 등을 고려할 수밖에 없다.

본 논문은 사용자의 최신 관심사를 반영하기 위해 사용자의 소셜 미디어의 활동을 바탕으로 사용자의 관심 카테고리에 시간 가중치를 부여함으로써 관심사의 최신성을 고려하고 오래된 관심사에 대한 가중치를 낮추었으며 유사한 사용자의 관심 카테고리를 고려하고 이를 반영하여 다양한 검색 결과를 제공한다.

Ⅲ. 제안하는 소셜 검색 기법

1. 전체적인 처리과정

기존 소셜 검색은 자신과 가까운 사람, 영향력이 큰 사람 그리고 공통 관심사를 이용한 검색 결과를 제공한다. 하지만 사용자의 최근 관심사와 유사한 사용자의 관심사를 고려하지 못했기 때문에 이들을 고려한 연구가 필요하다. 따라서, 본 논문에서는 소셜 미디어 사용자의 최근 관심사를 이용하고 유사한 사용자의 관심사를 이용한 소셜 검색 기법을 제안한다. 제안하는 소셜 검색의 특징은 사용자들의 소셜 활동 정보에 대해 텍스트 마이닝 기법을 사용하여 추출한 관심사를 카테고리화하고 사용자들의 프로파일을 생성하여 질의에서 추출한 관심사의 카테고리와 비교해 검색을 수행한다. 그리고 소셜 미디어 상에서 나와 성향이 비슷한 사용자, 특정 카테고리의 전문가 그리고 친구로 연결된 사용자들의 유사도를 고려하여 소셜 검색의 결과에 반영하였다. 또한 사용자들의 관심사들에 대한 시간적인 가중치를 부여하여 오래된 관심사의 비중을 감소시켰다.

제안하는 소셜 검색은 크게 사용자 프로파일 생성, 사용자 유사도 판별, 랭킹 알고리즘으로 나누어져있다. 사용자 프로파일 생성은 사용자의 소셜 활동 정보와 소셜 미디어에서 가져온 관심사로 사용자 프로파일을 생성하고 관심사에 대한 시간 가중치를 부여하여 최근 관심사만을 고려한다. 사용자 유사도 판별은 관심 카테고리가 유사한 사용자와 특정 카테고리의 전문 지식을 가진 사용자를 고려하여 판별한다. 랭킹 알고리즘은 질의

에 따른 웹 검색 결과에 사용자의 관심사를 적용시킨 후 유사한 사용자의 관심사를 적용시킨 검색 결과를 보여준다. [그림 1]은 제안하는 소셜 검색 기법의 처리 과정을 보여준다. 사용자 프로파일 생성과 사용자 유사도 판별의 과정은 전 처리로 수행 되고 랭킹 알고리즘은 질의가 들어왔을 때 수행이 된다. 사용자 프로파일 생성은 TF-IDF(Term Frequency - Inverse Document Frequency)기법을 이용하여 사용자의 관심사를 추출한다. 추출된 관심 카테고리에 시간 가중치를 적용하여 사용자의 최신 관심사를 고려한다. 사용자 유사도는 각각의 사용자 프로파일 정보를 활용하여 사용자와 관심 카테고리가 유사하고 특정 카테고리에 전문성 있는 사용자들을 판단하여 관심 카테고리 정보를 제공한다. 랭킹 알고리즘은 앞서 얻은 사용자의 관심 카테고리와 다른 사용자의 관심 카테고리를 이용하여 검색 결과를 재정렬한다.

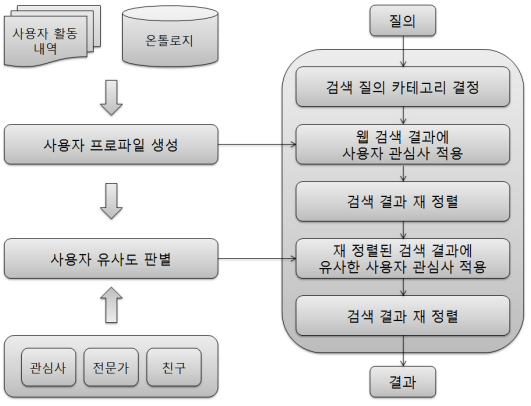

그림 1. 제안하는 소셜 검색 기법의 처리 과정

2. 사용자 프로파일 생성

사용자 프로파일은 사용자의 소셜 미디어 활동 내역을 통해 얻은 관심 키워드와 소셜 미디어의 페이지 분류를 위한 카테고리를 이용해 온톨로지화 하여 사용자의 관심사를 카테고리화 한다. 본 논문에서 제안하는 기법은 사용자 소셜 활동 카테고리를 이용하여 사용자의 관심사를 찾기 때문에 소셜 미디어 활동에서 얻은 관심사를 카테고리로 분류하는 과정이 필요하다. 카테고리를 이용한 사용자의 관심사 추출은 사용자가 생성하거나 공유한 데이터를 기반으로 한다. 사용자가 데이터를 생성하거나 공유, 댓글 등의 행동을 하는 것은 사용자의 관심사를 표출하는 것으로 해석할 수 있다.

[그림 2]는 사용자 프로파일 생성 과정을 나타낸다. 소셜 미디어로부터 사용자의 소셜 활동 데이터를 받아온 후 TF-IDF를 이용하여 소셜 활동에서 많이 언급되고 중요한 키워드들을 추출하게 된다. TF는 단어빈도로 특정 단어가 문서 내에 얼마나 많은 빈도로 등장하는지를 나타내고 IDF는 역문서 빈도로 전체 문서들에서 특정 단어를 가지고 있는 문서가 얼마나 되는지를 말한다. 예를 들어 사용자가 활동한 전체 문서가 1000개라고 가정하고 그중 '리버풀'이라는 단어가 나오는 문서는 3개라고 가정한다. 그 중 한 문서에서 '리버풀'이라는 단어가 3번 나왔다고 가정하게 되면 그 문서에 대한 TF는 3 IDF는 2.5215가 된다. 즉, TF-IDF는 7.5645라는 값을 가지게 되며 점수가 높을수록 중요한 키워드가 된다. 이렇게 추출된 키워드들의 카테고리 분류는 소셜 미디어에 존재하는 페이지 분류를 위한 카테고리를 사용하여 온톨로지를 구축하였고[14] OWL[15]을 이용하여 추출한 키워드들을 카테고리화 하였다. 이를 이용해 작성된 카테고리는 사용자의 프로파일이 되며 사용자의 현재 관심사가 된다.

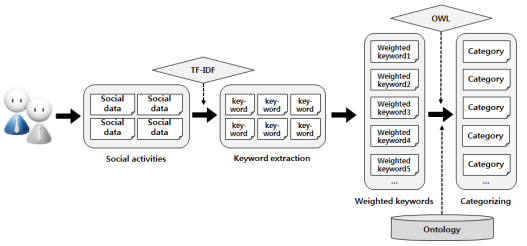

그림 2. 사용자 프로파일 생성 과정

식 (1)은 관심 카테고리의 점수를 구하는 방법을 기술한 식이다. 관심 카테고리안의 키워드의 빈도수가 높다는 것과 가장 최근에 검색한 키워드는 키워드에 대한 관심이 높다는 것이다. 그에 따라 사용자의 관심 카테고리를 시간에 따른 가중치를 부여하였고 평균의 맹점을 최소화하기 위해 키워드 빈도수에 지수를 취하였다. 추출된 키워드를 카테고리화한 관심사의 점수 TW는

관심 키워드에 대한빈도수와 시간에 따른 가중치로 표현을 하였다. n은 관심사에 대한 키워드 수, $tw_i = (-1.0506^t + 2)$는 시간에 따른 키워드의 가중치, cnt_i는 키워드의 빈도수를 나타낸다.

$$TW = \frac{1}{n}\sum_{i=1}^{n} tw_i * e^{cnt_i} \quad (1)$$

기존에 제안된 사용자 프로파일 기법들은 사용자의 활동내역과 관심분야의 분석을 실시하지 않는 비실시간적인 프로파일 정보를 생성하여 소셜 검색을 수행한다. 대부분의 소셜 미디어에서는 사용자의 프로파일을 주기적으로 갱신하지 않기 때문에 사용자의 최근 관심사를 파악하기 어렵다. 따라서 본 논문에서는 오래된 관심사의 비중을 감소시키고 최근 관심사에 대한 비중을 높이기 위해 오래된 관심사와 최근 관심사에 대하여 시간에 따른 가중치를 부여하였다. 그리고 현재 관심사에 대한 프로파일을 구축하기 위해 최근 활동을 분석하고 오래된 관심사의 비중을 감소시키는 프로파일을 제안한다. 사용자의 관심사를 파악하는 것은 사용자의 특성을 분석 하는 데에 있어서 중요한 요소이다. 사용자의 관심 키워드를 추출하고 카테고리화한 사용자의 특성은 소셜 검색을 제공하는 것에 있어서 매우 중요한 사항이 된다. 그리고 개개인의 사용자들은 각각의 관심사에 대해서 각기 다른 관심 비중을 가지고 있다. 따라서 본 논문에서는 카테고리화한 관심사들에 사용자들의 특성을 부여하고자 관심사 가중치를 부여하였다.

3. 사용자 유사도 판별

개인화된 랭킹을 적용하는 일에는 여러 가지 위험이 따른다. 평소에 A라는 문서를 선호하는 사용자라도 가끔 B 문서에 대해 관심을 가질 때가 있을 것이다. 만약 A라는 문서만 제공하면 오히려 사용자 만족도를 떨어뜨리는 결과를 초래한다. 이와 마찬가지로 사용자의 관심사를 파악하는 것도 중요하지만 사용자의 특성만을 분석하여 개인화 검색을 하게 되면 사용자의 특성만을 반영한 편협한 정보만이 검색 결과로 도출된다[10].

사용자의 지인, 관심사가 비슷한 사용자, 특정 관심사의 전문가 등 사용자와 연관이 되어있는 대상을 고려하여 검색하면 더 신뢰성이 있고 사용자가 원할만한 데이터를 검색결과로 보여줄 수 있다. 따라서 사용자와 유사한 사용자를 판별하기 위해서 관심 카테고리가 유사한 사용자와 특정 카테고리의 전문가 그리고 친구로 연결된 사용자의 관심사를 고려한다. 사용자와 유사한 성향을 가진 사용자들, 관심사에 대한 활동이 많은 사용자들 그리고 사용자와 소셜 미디어에서 사용자와 소셜 관계가 높은 사용자들의 성향을 고려하여 사용자에게 검색 정보를 제공한다.

식 (2)는 사용자와 유사한 성향을 가진 사용자들에 대한 점수를 구하는 식이다. 관심 카테고리 점수의 차가 적은 것은 관심 카테고리에 대한 관심도가 비슷하다는 것을 의미한다. 그래서 사용자의 관심 카테고리 점수와 다른 사용자의 관심 카테고리 점수의 차를 구해 그 차이 적은 사용자들을 관심 카테고리가 유사한 사용자라고 정의하였다. 그리고 관심이 일치하는 카테고리를 활용하여 사용자의 관심 카테고리의 수로 나누었다. n은 사용자의 관심 카테고리 수, k는 사용자와 다른 사용자와 일치하는 관심 카테고리 수이다.

$$S_1 = \frac{1}{n}\sum_{i=1}^{k}(1 - |TW_{ui} - TW_{oi}|) \quad (2)$$

식 (3)은 사용자의 관심 카테고리에 대한 활동이 많은 사용자들에 대한 점수를 구하는 식이다. 특정 문서에 대한 다른 사람들의 참여도가 높을수록 문서를 작성한 사용자의 전문성이 높다는 것을 의미한다. 사용자의 소셜 미디어 활동에 대해 사람들의 게시물 참여자 수에 따른 추천, 공유, 코멘트를 고려하여 전문성 있는 유서를 식별하였다. n은 다른 사용자의 활동 페이지의 수 N은 페이지에 대한 참여자의 수, NR은 추천 수, NS는 공유 수, NS는 코멘트 수이다.

$$S_2 = \frac{1}{n}\sum_{i=1}^{n}\frac{NR_i + NS_i + NC_i}{N_i} \quad (3)$$

식 (4)는 소셜 미디어에서 사용자와 소셜 관계가 높은 사용자들의 점수를 구하는 식이다. 자신과 가까운 사용자는 자신에게 만족할만한 정보를 제공할 수 있기 때문에 '케빈 베이컨의 6단계 법칙'을 적용하여 소셜 미디어에서 사용자와 가까운 홉일수록 높은 가중치를 부

여하였고 7홉이 넘어가면 가중치가 부여되지 않는다.

$$S_3 = e^{1-\frac{1}{6}x} - 1 \quad (4)$$

식 (5)는 앞에서 고려한 (2), (3), (4)번의 식에 각각 가중치를 두어 사용자 유사도를 고려하였다. 세 가지의 수식을 고려한 사용자 유사도를 가지고 사용자 프로파일에서 사용자와 유사한 사용자를 판별하게 된다. 사용자의 관심사, 유사한 사용자의 관심사, 친구로 연결된 사용자 수식에 가중치를 적절히 부여하여 어떤 요소가 사용자에 더 큰 영향을 미치는지 확인해 볼 수 있다. 가중치 α, β, γ의 합은 1로 표현된다.

$$USIM = \alpha S_1 + \beta S_2 + \gamma S_3 \quad (5)$$

식 (5)를 이용하여 사용자 유사도를 구해보도록 한다. [표 1]은 사용자들에 대한 관심 카테고리와 소셜 활동을 나타내고 있다. 사용자 A를 기준으로 사용자 B와 C의 사용자 유사도를 구해보도록 한다. 먼저 사용자 유사도는 관심 카테고리 점수 TW를 이용하여 S_1을 구할 수 있다. 사용자 A와 B의 S_1 점수는 0.5점이 되고 사용자 A와 C의 S_1 점수는 0.5점이 된다. 그다음 특정 카테고리의 전문가는 소셜 활동 정보를 이용하여 S_2를 구할 수 있다. 수식 (3)을 이용하면 사용자 B와 C의 S_2 점수는 각각 0.42, 0.7점이 된다. 그리고 사용자 A, B, C의 소셜 관계가 각각 1홉이고 가중치 α, β, γ의 값을 각각 0.45, 0.45, 0.1이라고 가정하면 사용자 A와 B 그리고 A와 C의 유사도 점수는 각각 0.514, 0.64점이 된다. 따라서 사용자 A와 가장 유사한 사용자는 사용자 C가 된다.

표 1. 사용자의 관심 카테고리와 소셜 활동

사용자	관심 카테고리 (TW점수)	키워드	소셜 활동
A	노래(0.5)	어쿠스틱 콜라보, 슈가볼	노래 (좋아요 : 50, 댓글 : 50, 공유 : 5, 참여자수 500)
	스포츠 리그(0.5)	프리미어리그, 프리메리리가	
B	스포츠 리그(0.5)	프리미어리그, 프리메리리가	스포츠 리그 (좋아요 : 100, 댓글 : 100, 공유 : 10, 참여자수 500)
	스포츠 팀(0.5)	리버풀, 바르셀로나	
C	노래(0.5)	어쿠스윗, 랄라스윗	노래 (좋아요 : 80, 댓글 : 60, 공유 : 10, 참여자수 200)
	뮤직 비디오(0.5)	러시안룰렛, cheer up	

사용자의 관심 키워드 외에 다른 사용자들의 관심 키워드들을 추천 받기 위해 사용자와 유사한 사용자들의 키워드들 고려하였다. 제안하는 기법에서 사용자의 관심사와 나와 비슷한 성향을 가진 다른 사용자들의 관심사를 파악할 수 있다. 이를 이용하여 사용자가 관심 있어 하는 카테고리 안에서 다른 키워드를 포함한 검색결과를 받아볼 수 있다. 앞서 3장에 언급한 것처럼 다른 사용자의 관심사를 받아보는 것은 중요한 일이다. 하지만 소셜 검색은 다른 사용자들의 관심사들 보다 사용자의 관심사가 더욱 중요하다. 사용자의 관심사를 다른 유사한 사용자의 관심사 보다 비중이 더 높아야한다. 그래서 랭킹 알고리즘의 랭킹은 사용자의 관심사를 다른 사용자의 관심사 보다 높은 우선순위를 가지기로 한다.

4. 랭킹 알고리즘

검색의 질을 결정하는 데에 있어서 가장 중요한 요소가 랭킹 알고리즘이다. 사용자들은 더 정확하고, 자신이 원하는 정보를 검색하길 원한다. 그래서 사용자의 관심사가 반영된 소셜 검색 결과를 보여주기 위해 사용자의 관심사, 사용자와 유사한 사용자, 특정 카테고리의 전문가 그리고 친구로 연결된 사용자들의 관심사를 종합하여 순위를 결정한다.

[그림 3]은 랭킹 알고리즘의 의사코드(pseudo code)를 보여준다. 먼저 사용자의 검색 질의로부터 키워드를 추출하여 검색의 카테고리를 결정한다. 그리고 사용자 질의에 대한 웹 검색 결과를 받아온다. 그 후 카테고리의 키워드들을 fractional-cascading 방식[16]으로 매칭시키고 사용자의 관심 키워드의 일치 수가 많을수록 상위랭크로 올린다. 그 후에 유사한 사용자의 관심 키워드를 fractional-cascading 방식으로 매칭시켜 다시 한 번 랭킹을 결정하고 웹 검색 결과를 재 정렬한다. 랭킹 알고리즘에서 사용자의 관심 카테고리를 고려한 후에 유사한 사용자의 관심사를 적용하였는데 이는 다른 사용자들의 관심 키워드에 대한 제약이다. 랭킹 알고리즘에서 키워드 매칭 시에 다른 사용자들의 관심 키워드의 경우 사용자의 관심 키워드와 반드시 같이 동반되어 검색 결과에 반영되어야한다. 그 이유는 사용자의 관심사를 다른 유사한 사용자의 관심사 보다 비중이 더 높아

야하며 다른 사용자의 관심 키워드가 사용자의 선호 키워드가 아닌 경우 사용자에게 적합하지 않은 정보를 제공할 수 있기 때문이다.

알고리즘 1. Fractional-cascading 방식의 랭킹 알고리즘

```
Counter [1. . . D] <- 0

For each Web search result R
  For each the user interest keyword K∈D
    if (K matches R)
      counter[K] ++
    else
      break

return ranked Web search result R'

For each ranked Web search result R'
  For each other user interest keyword K'' ∈D
    if (K'' matches R')
      counter[K''] ++
    else
      break

return ranked Web search result R''
```

그림 3. 랭킹 알고리즘

Ⅳ. 성능평가

본 논문에서 소셜 미디어를 통하여 사용자들의 관심사를 판단하고 이를 바탕으로 임의의 데이터로 1000개의 사용자 프로파일을 구축한다. 그리고 특정 사용자를 선별해 사용자의 관심사와 유사한 사용자의 관심사를 실제 웹 검색에 매칭시켜 결과를 재 정렬하여 검색 성능과 만족도가 얼마나 반영 되었는지를 확인한다. 실험 환경은 Intel core i5-4460 CPU 3.20GHz, 8GB의 메모리를 갖는 시스템에서 JAVA로 구현하였다. 본 연구에서는 실제 웹 검색 결과를 사용하여 사용자의 관심사를 적용하여 재 정렬하고 검색의 정확도와 만족도를 평가하는 실험을 하였다.

[표 2]은 본 논문에서 제시하는 기법이 기존 검색 결과보다 검색 결과를 얼마나 효율적으로 제공하는지를 측정하기 위해 상위에 제시된 결과들을 이용하여 새로운 랭킹리스트를 생성했다. 왼쪽에 있는 결과는 검색엔진 구글의 검색 결과이며 오른쪽에 있는 결과는 제안하는 소셜 검색의 검색 결과이다. 이 결과는 기존의 결과 값보다 사용자의 개인 성향을 반영한 결과 값을 상위에 노출시켰다.

표 2. 특정 질의에 대한 검색 결과

질의(프리미어 리그)에 대한 검색 결과	
기존 웹 검색 기법	제안하는 기법
프리미어리그 - 위키백과, 우리 모두의 백과사전	T.P.T.P - LIVERPOOL FC KOREAN FAN SITE
잉글리시 프리미어 리그 - 나무위키	[2015/16시즌] 프리미어리그에서 교체를 가장 잘 하는 감독 클롭...
축구: 프리미어리그 2015/2016 실시간 -결과, 일정, 순위 ...	한국일보 : [듀어든] 손흥민, 왜 하필 토트넘에 갔을까
잉글랜드 프리미어리그 2015-2016 _데이터_모든 경기일정 ...	11/12 잉글리쉬 프리미어리그 2R - 아스널 vs 리버풀 110820 ...
해외축구 EPL 경기 일정/결과	다음스포츠 - Daum	'프리미어리그' 토트넘 웨스트브롬 1-1 무승부…손흥민 후반 ...
프리미어 리그 순위 - Goal.com	스포츠 · 연예 : [프리미어리그] 손흥민 `후보`...토트넘 VS 웨스트 ...
영국 서민은 왜 한달에 한번도 축구 경기를 볼 수 없나 : 유럽 ...	[프리미어리그] 손흥민 출전 토트넘, 웨스트브롬과 무승부 ...
잉글리시 프리미어리그 : SBS Sports	손흥민 후반 교체투입′ 토트넘, WBA와 1-1... 우승 희망 멀어져

본 논문에서는 제안한 랭킹 알고리즘과 기존의 웹 검색의 알고리즘을 비교하기 위해 기존의 웹 검색 방식과 제안하는 기법이 적용된 검색 결과의 만족도를 평가하는 nDCG(Normalized Discounted Cumulative Gain)와 정확도의 성능척도인 정확도(precision)와 재현율(recall) 이용하여 두 가지 방법으로 평가한다. 정확도와 재현율은 정보검색에서 중요한 성능 측정 기준으로 사용하는 지표이다. 정확도는 정보 검색 분야에서 검색된 문서들 중 관련 있는 문서들의 비율이고 재현율은 정보 검색 분야에서 관련 있는 문서들 중 실제로 검색된 문서들의 비율이다. 정확도는 수식(6)과 같이 전체 검색 결과에 대해 질의와 관련 있는 문서 개수를 나타내는 비율이며, 재현율은 수식(7)과 같이 문서 개수를 기준으로 했을 때 문서 개수에서 관련 있는 문서를 나타내는 비율이다.

$$Precision = \frac{Relevant\ \ D \cap Retrieved\ \ D}{Retrieved\ \ D} \qquad (6)$$

$$Recall = \frac{Relevant\ \ D \cap Retrieved\ \ D}{Relevant\ \ D} \qquad (7)$$

검색 결과에 대한 사용자의 만족도를 측정하기 위해 정보검색에서 만족도 측정으로 사용되는 nDCG(normalizing Discounted Cumulative Gain)@k를 활용한다. nDCG는 기존의 정확도, 재현율 기반의 검색엔진 평가 방법으로는 순위에 따른 차별점을 부여하기 어렵다는 판단에 따라 나온 방법이며 사용자들이 원하는 문서를 상위에 올려줄수록 만족도가 높아지는 평가 방식이다. 수식 (8)은 검색 결과의 문서 각각에 대한 만족도 점수를 상위에 노출된 순으로 가중치를 두어 검색 결과에 대한 만족도를 나타낸다. 수식 (9)에서 IDCG는 사용자의 검색 만족도의 이상적인 점수다. nDCG점수는 DCG와 IDCG의 비율을 통해 구할 수 있다.

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i} \quad (8)$$

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (9)$$

정확도와 재현율은 [그림 4]와 [그림 5]로 나타낼 수 있다. 각 그림은 검색결과의 20단위의 누적 값과 분할 값을 가지고 있다. 기준을 상위 40개로 가정하였을 때 정확도는 기존보다 177% 향상이 되었고 상위 20개의 재현율은 163%가량 향상된 것을 확인할 수 있다. 정확도의 경우 전 구간에서 기존 검색 엔진보다 제안하는 기법이 우수하다는 것을 알 수 있다. 재현율의 경우 상위 40개 구간에서 기존 검색 엔진보다 사용자가 원하는 정보를 제공한 것을 알 수 있다. 이는 제안하는 알고리즘이 사용자에게 원하는 정보를 상위로 노출 시킨 것을 알 수 있다.

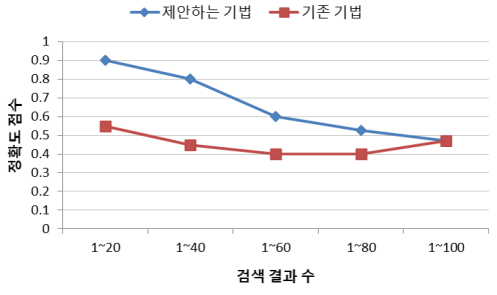

그림 4. 정확도

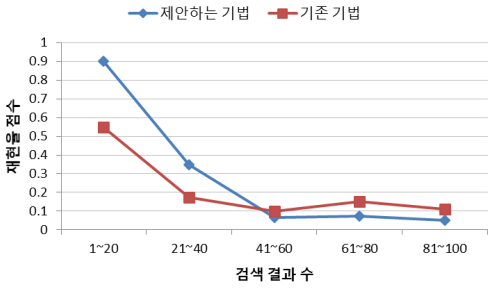

그림 5. 재현율

[그림 6]은 '프리미어 리그'라는 질의의 결과 100개에 대한 네 가지 경우의 nDCG점수를 나타낸 수치다. [그림 6]을 통해 기존의 웹 검색 보다 본 논문에서 제안한 알고리즘을 적용한 검색결과가 더 좋다는 것을 알 수 있었다. 그리고 사용자의 관심사, 유사한사용자의 관심사, 친구로 연결된 사용자의 가중치인 α, β, γ의 비중을 다르게 두어 성능평가를 실시하였다. case1의 경우 α=0.33, β=0.33, γ=0.33, case2의 경우α=0.4, β=0.4, γ=0.2, case3의 경우α=0.45, β=0.45, γ=0.1의 가중치를 두었다. 이 결과를 통해 관심 카테고리가 비슷한 사용자와 관심 카테고리에 대한 활동이 많은 사용자의 비중이 소셜 관계가 높은 사용자들 보다 검색결과에 더 큰 영향을 미치는 것을 알 수 있다.

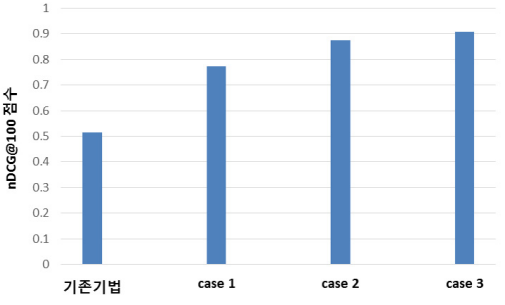

그림 6. nDCG@100 점수

V. 결 론

본 논문에 제시한 방법은 사용자의 최근 성향을 반영하는 검색결과를 효과적으로 상위에 배치시키는 장점

이 있다. 특정 질의로 검색한 웹 검색 결과를 사용자가 원하는 검색 결과로 재정렬시켜 사용자에 맞는 검색 결과를 향상시켰다. 제안하는 소셜 검색은 웹 검색을 대상으로 하기 때문에 높은 활용도가 예상되며 사용자의 최근 성향을 반영하여 오래된 관심 카테고리를 반영하지 않을 수 있는 장점이 있다. 성능 평가 결과 상위 40개의 검색 결과는 기존보다 정확도가 177% 향상 되었고 상위 20개의 검색 결과의 재현율은 163% 향상된 것을 확인할 수 있었다. 정확도의 경우 전 구간에서 기존보다 제안하는 기법이 우수하다는 것을 알 수 있으며 재현율의 경우 사용자의 성향의 반영된 검색 결과가 상위 40개의 검색 결과에 많이 분포한 것을 알 수 있었다. 또한 검색 결과에 대한 사용자의 만족도를 평가하여 기존보다 제안하는 기법의 우수성을 보였으며 사용자의 관심사, 유사한 사용자의 관심사, 친구로 연결된 사용자의 가중치인 α, β, γ의 비중에 따른 만족도가 사용자의 관심사와 유사한 사용자의 관심사 비중이 소셜 관계가 높은 사용자들 보다 검색결과에 더 큰 영향을 미치는 것을 알 수 있다. 향후 연구로 실험 결과의 양적인 면과 질적인 면을 발전시켜 다양한 실험을 할 필요가 있다. 또한, 추후에는 사용자들의 질의에 대한 검색결과 만족도를 피드백을 추가하여 검색 결과의 정확도를 더욱 향상시킬 수 있는 방법을 찾아내야 하며 검색결과에 대해 사용자의 만족도를 어떠한 방식으로 분석해야할 것인지는 추후에 연구해야할 과제이다.

참 고 문 헌

[1] M. J. Morris, “Teevan and K. Panovich, “What Do People Ask Their Social Networks, and Why?,” Proc. International Conference on Human Factors in Computing Systems, pp.1739-1748, 2010.

[2] 김주영, 조찬형, 장세정, 윤은정, “*2015년 모바일인터넷이용실태조사 최종보고서*,” 한국인터넷 진흥원, 2015.

[3] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu, “Exploring folksonomy for personalized search,” Proc. International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.155-162, 2008.

[4] M. Claypool, P. Le, M. Waseda, and D. Brown, “Implicit interest indicators,” Proc. International Conference on Intelligent User Interfaces, pp.33-40, 2001.

[5] B. Carterette and R. Jones, “Evaluating Web Search Engines Using Clickthrough Data,” Proc. International ACM SIGIR Conference, 2007.

[6] D. Horowitz and D.Kamvar, “The Anatomy of a LargeScale Social Search Engine,” Proc. International Conference on World Wide Web, pp.431-440, 2010.

[7] S. Bao, G. Xue, X. Wu, Y.Yu, B. Fei, and Z.Su, “Optimizing Web Search Using Social Annotations,” Proc. International Conference on World Wide Web, pp.501-510, 2007.

[8] A. Kashyap, R. Amini, and V. Hristidis, “SonetRank: Leveraging Social Networks to Personalize Search,” Proc. ACM International Conference on Information and knowledge management, pp.2045-2049, 2012.

[9] O. Shafiq, R. Alhajj, and John G. Rokne, “On personalizing Web search using social network analysis,” International Journal of Information Sciences, Vol.314, pp.55-76, 2015.

[10] K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag, “Personalizing Web Search Results by Reading Level,” Proc. ACM International Conference on Information and knowledge management, pp.403-412, 2011.

[11] 이동균, 권준희, “최근 사용자 관심사를 고려한 소셜 검색 알고리즘,” 한국정보기술학회논문지, 제9권, 제4호, pp.187-194, 2011.

[12] Y. A. Kim and G. W. Park, “Topic-Driven SocialRank: Personalized search result ranking

by identifying similar, credible users in a social network," International Journal of Knowledge-Based Systems, Vol.54, pp.230-242, 2013.

[13] 복경수, 안민제, 임종태, 유재수, "모바일 환경에서 시간 속성을 고려한 효율적인 위치 기반 소셜 검색," 정보과학회논문지, 제20권, 제4호, pp.243-247, 2014.

[14] T. Berners-Lee, J. Hendler, and O. Lassila, *The Semantic Web*, Scientific American Magazine, 2001.

[15] M. Dean and G. Schreiber, *OWL Web Ontology Language*, 2004.

[16] B. Chazelle and L. J. Guibas, "Fractional cascading: I. A data structuring technique," International Journal of Algorithmica, Vol.1, No.2, pp.133-162, 1986.

저 자 소 개

송 진 우(Jinwoo Song) 준회원

- 2015년 2월 : 충북대학교 정보통신공학과(공학사)
- 2015년 3월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : 소셜 네트워크, 소셜 검색, 빅데이터

전 현 욱(Hyeonwook Jeon) 준회원

- 2015년 2월 : 강원대학교 정보통신공학과(공학사)
- 2015년 3월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : 빅데이터, 스트리밍, 스케줄링

김 민 수(Minsoo Kim) 준회원

- 2013년 2월 : 충북대학교 정보통신공학과(공학사)
- 2013년 ~ 2014년 : (주) 매크로 임팩트 연구원
- 2015년 3월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : 빅데이터, RDF, 분산처리

김 기 훈(Gihoon Kim) 준회원

- 2016년 2월 : 충북대학교 정보통신공학과 (공학사)
- 2016년 3월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : 빅데이터, 고차원 인덱스

노 연 우(Yeonwoo Noh) 준회원

- 2014년 2월 : 충북대학교 정보통신공학과(공학사)
- 2016년 2월 : 충북대학교 정보통신공학과(공학석사)
- 2016년 3월 ~ 현재 : 충북대학교 정보통신공학과 박사과정

<관심분야> : 데이터베이스시스템, 소셜 네트워크 서비스, 빅데이터 등

임 종 태(Jongtae Lim) 정회원

- 2009년 2월 : 충북대학교 정보통신공학과(공학사)
- 2011년 2월 : 충북대학교 정보통신공학과(공학석사)
- 2015년 8월 : 충북대학교 정보통신공학과(공학박사)
- 2015년 9월 ~ 현재 : 충북대학교 정보통신공학과 박사후연구원(Post.doc)

<관심분야> : 데이터베이스 시스템, 시공간 데이터베이스, 위치기반 서비스, 모바일 P2P 네트워크, 빅데이터 등

복 경 수(Kyungsoo Bok) 종신회원

- 1998년 2월 : 충북대학교 수학과 (이학사)
- 2000년 2월 : 충북대학교 정보통신공학과(공학석사)
- 2005년 2월 : 충북대학교 정보통신공학과(공학박사)
- 2005년 3월 ~ 2008년 2월 : 한국과학기술원 전산학과 Postdoc
- 2008년 3월 ~ 2011년 2월 : ㈜가인정보기술 연구소
- 2011년 3월 ~ 현재 : 충북대학교 정보통신공학과 초빙교수

<관심분야> : 데이터베이스 시스템, 위치기반서비스, 모바일 P2P 네트워크, 소셜 네트워크 서비스, 빅데이터 등

유 재 수(Jaesoo Yoo) 종신회원

- 1989년 2월 : 전북대학교 컴퓨터공학과(공학사)
- 1991년 2월 : 한국과학기술원 전산학과(공학석사)
- 1995년 2월 : 한국과학기술원 전산학과(공학박사)
- 1995년 2월 ~ 1996년 8월 : 목포대학교 전산통계학과 전임강사
- 1996년 8월 ~ 현재 : 충북대학교 전자정보대학 정교수

<관심분야> : 데이터베이스 시스템, 빅데이터, 센서네트워크 및 RFID, 소셜 네트워크 서비스, 분산 객체컴퓨팅, 바이오인포매틱스 등