디지털 도서관 이용자의 검색행태 연구*

- 검색 로그 데이터의 네트워크 분석을 중심으로 -

A Study on the Search Behavior of Digital Library Users:

Focus on the Network Analysis of Search Log Data

이 수 상(Soo-Sang Lee)**
위 성 광(Cheng-Guang Wei)***

〈 목 차 〉

초 록

본 논문에서는 검색로그 데이터의 네트워크 분석방법을 통해 검색자들의 검색행위에 나타난 다양한 특성을 살펴보았다. 이러한 작업을 통해 얻어진 결과는 다음과 같다. 첫째, 검색자들은 검색어의 유사성에 따라 네트워크라는 연결구조를 나타내었다. 둘째, 특정한 검색자 네트워크에서 중심적인 위치를 차지하는 검색자들이 존재하였다. 셋째, 중심 검색자들은 다른 검색자들과 검색 키워드를 공유하고 있었다. 넷째, 전체 검색자들은 다수의 하위 집단으로 군집되어 있다. 이 연구의 결과는 네트워크 분석 방법에 의한 연관된 검색자와 검색어를 추천하는 알고리즘을 개발하는데 활용이 가능할 것이다.

키워드: 검색로그, 네트워크 분석, 중심성 분석, 에고 네트워크, 군집 분석

ABSTRACT

This paper used the network analysis method to analyse a variety of attributes of searcher's search behaviors which was appeared on search access log data. The results of this research are as follows. First, the structure of network represented depending on the similarity of the query that user had inputed. Second, we can find out the particular searchers who occupied in the central position in the network. Third, it showed that some query were shared with ego-searcher and alter searchers. Fourth, the total number of searchers can be divided into some sub-groups through the clustering analysis. The study reveals a new recommendation algorithm of associated searchers and search query through the social network analysis, and it will be capable of utilization.

Keywords: Search Log, Network Analysis, Centrality Analysis, Ego Network, Clustering Analysis

* 이 논문은 2009년 5월 29일 한국도서관·정보학회 하계학술발표대회에서 발표했던 내용을 수정·보완한 것임.
이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음.
** 부산대학교 문헌정보학과 부교수(sslee@pusan.ac.kr)(제1저자)
*** 부산대학교 대학원 박사과정(weipnu@hanmail.net)(공동저자)
• 접수일: 2009년 11월 20일 • 최초심사일: 2009년 11월 30일 • 최종심사일: 2009년 12월 26일

Ⅰ. 서 론

일반적으로 로그 분석(log analysis)은 웹서버에 사용자가 들어오는 순간부터 하나의 데이터에 접속(hit), 실제 이용자가 하나의 완성된 페이지를 보는 행위(view), 특정 사용자가 일정시간 내에 계속적으로 웹서버를 검색(search)하는 등 웹서버의 방문(visit) 데이터를 기반으로 어떤 목적에 맞도록 분석을 수행하는 계량적 방법을 말한다. 이와 같은 다양한 방문 데이터들이 통계분석의 대상이 될 수 있으며, 이를 바탕으로 해당 기관의 웹서버에 대하여 얼마나 많은 사람들이, 언제 방문하는지, 가장 오래 보는 자료와 가장 많이 보는 자료는 어떤 것인지 등 다양하고 의미 있는 정보들을 파악해 낼 수 있다.[1)]

로그분석에 사용되는 로그파일은 이용자와 웹 서버간의 상호작용과정에서 생성되는 데이터로 문자와 숫자의 조합으로 이루어진다. 로그분석을 하는 목적은 이용자가 남긴 정확한 로그데이터를 바탕으로 이용자의 행태를 분석하여 이용자에게 유용한 웹서비스를 구축하려는데 있다. 즉, 이용자들이 자주 찾는 콘텐트, 주로 입력하는 키워드 등을 토대로 콘텐트나 키워드 등을 추천할 수 있다. 그리고 이용자들의 접속 경로나 네비게이션 행태를 파악하여 웹 사이트의 구조를 조정할 수 있다. 이러한 유용성에도 불구하고, 이용자 개인의 프라이버시를 침해할 수 있는 문제, 서비스 이용에 대한 질적인 판단과 분석의 문제, 로그데이터 자체의 정확성 여부 판단이 불확실한 경우 등과 같은 단점도 있다. 아무튼 로그분석의 결과는 로그데이터에 나타난 바대로 이용자의 행태로 인식하는 것이 중요하며, 그것에 너무 과도한 의미를 부여하는 데는 무리가 있을 수 있다.

본 논문에서는 로그데이터 특히 검색과정에서 얻어진 검색로그 데이터를 대상으로 하는 검색로그 분석을 시도하려고 한다. 검색로그 분석은 이용자가 해당 웹 사이트를 방문하여 검색, 클릭, 이동 등과 관계된 검색 로그 파일만을 분석하여 이용자의 실제적인 검색 행위를 파악하는 분석기법이다.[2)] 이용자들이 검색이라는 트랜잭션을 통해 생성한 데이터로부터 유용한 정보를 인지하고, 그것을 분석하여 특정한 의사결정을 위한 지식으로 활용할 수 있게 된다.

검색로그 분석은 크게 검색 과정에서 나타나는 질의로그(query log)의 분석과 클릭로그(click log)의 분석으로 구분할 수 있다. 그리고 검색로그 분석을 통해 이용자들의 검색행위에 나타난 다양한 특성들을 파악할 수 있는데, 이 중에서도 질의에 나타난 검색어의 특성, 클릭에 나타난 콘텐트의 선택 등으로 연관된 검색어, 검색자, 콘텐트 등을 다른 검색자들에게 추천할 수 있다. 따라서 본 논문은 이러한 검색로그 분석의 다양한 측면을 모두 고려하기 보다는 시론적 차원에서 질의로그에 나타난 검색어를 토대로 검색자의 검색행태를 다양한 관점에서 파악하고자 한다.

---

1) 서진완, "로그화일(Log file)을 이용한 공공기관의 홈페이지 분석과 정책적 함의," 한국행정학회, 춘계학술대회 발표논문집(2001), pp.501-517.

2) 박소연·이준호, "웹 검색 분야에서의 로그 분석 방법론의 활용도," 한국문헌정보학회, 학술발표논집, 제21집(2006), pp.81-94.

검색어는 검색 이용자가 검색 시점에 관심을 가진 주제를 표현한 것이라고 한다면, 유사한 관심 주제를 가진 검색자들을 연결하고 구분해 보는 작업은 의미가 있는 일이다. 검색자들의 연결구조에 나타난 다양한 특성은 검색자들의 검색행태를 설명할 수 있는 근거가 되며, 그것은 또한 정보검색 영역에 유용하게 활용할 수 있을 것이다. 이러한 검색자의 연결구조는 네트워크 분석방법에서 개발된 다양한 기법으로 파악이 가능하다. 즉, 검색로그에 나타난 검색어들을 통해 검색자들의 네트워크를 생성하고, 다양한 분석기법을 통해 검색자들의 검색행위의 특성들을 있는 그대로 인식하고, 그것에 내재된 의미들을 파악하고자 한다. 연관검색어를 추천하기 위하여 검색로그 데이터에 나타난 검색어들의 특성을 마이닝 기법을 사용하여 파악하는 연구는 있었지만, 검색자들의 관심 주제에 따른 사회적 연결구조를 파악하고자 한 연구는 거의 없었다. 그러므로 본 논문은 검색어의 특성보다 검색자들의 검색 행태로 나타나는 다양한 특성을 네트워크 분석기법으로 파악하고자 한다. 검색자들의 네트워크를 생성하고, 네트워크의 구조적 속성을 알아내고, 검색자들의 집단을 군집분석하며, 특정 집단 내의 검색자와 검색어의 특성을 분석하여 이용자의 다양한 검색 행태를 파악하게 된다.

Ⅱ. 검색로그 분석과 네트워크 분석

1. 검색로그 분석

검색로그 분석은 이용자가 해당 웹사이트를 통해 수행한 실제의 검색 행위와 관련한 로그데이터를 대상으로 이용자의 검색 행위를 정량적으로 분석하는 방법의 하나이다. 이러한 분석을 통해서 이용자의 검색 행위에 대한 다양한 정보를 얻게 되는데, 검색 로그분석의 목적을 구체적으로 살펴보면 크게 네 가지로 나눌 수 있다.

첫째, 이용자들의 검색 행위에 나타난 다양한 정보를 파악할 수 있다. 이용자들이 입력하는 검색 키워드는 곧 이용자들의 관심사를 나타내는 것이므로 이용자들이 많이 입력하는 검색 키워드들의 시간별, 계절별 변화를 분석하여 서비스의 개선에 활용이 가능하다. 둘째, 검색 키워드를 바탕으로 연관 검색어를 추천하고, 이용자는 이것을 활용하여 검색질의를 확장하도록 할 수 있다. 셋째, 개별 이용자에 적합한 서비스를 제공할 수 있다. 검색 로그파일에는 로그인 정보, IP주소, Cookie 등을 활용하여 개별 이용자를 식별하고, 이들의 검색행태와 관심사 등을 인지하여 개인별 맞춤형 서비스를 제공할 수 있다. 넷째, 검색로그 분석을 통해 웹사이트의 검색 성능을 향상시킬 수 있다. 이용자들이 많이 입력하는 검색 키워드를 바탕으로 시소러스를 작성할 수 있으며, 오타의 분석과 수정도 가능하여 검색 성능을 향상시킬 수 있다.

검색로그 분석을 분석에 사용한 검색로그의 유형에 따라 크게 두 가지로 구분할 수 있다. 첫 번째는 이용자가 검색을 위해 입력한 질의인 검색어만을 대상으로 분석하는 질의로그 분석 또는 검색어 로그 분석이다. 두 번째는 이용자가 입력한 검색어뿐만 아니라 검색 결과 중에서 이용자가 실제로 사용하기 위해 자료를 선택한 행위를 보여주는 클릭로그 데이터를 분석하는 클릭로그 분석 또는 트랜잭션 로그 분석이다.

질의로그 분석은 앞에서 언급한 것과 같이 사이트의 이용자가 검색을 위해 검색창에 입력한 검색어만을 대상으로 분석하는 방법이다. 주로 포털사이트를 대상으로 많은 연구가 이루어지고 있으며, 장기간에 걸친 방대한 자료를 바탕으로 이용자의 대략적인 검색 행태를 파악할 수 있다. 예를 들면, 알타비스타를 대상으로 한 연구에서는 이용자들은 검색을 위해 매우 짧은 검색어를 입력하고 한번 입력한 검색어는 거의 수정을 하지 않는 것으로 나타났다.[3] 그리고 네이버를 이용하는 이용자들의 검색 행태가 계절별, 요일별, 날짜별 그리고 주중과 주말에 따라 다르다고 한다.[4][5]

클릭로그 분석은 이용자가 실제로 관심을 가지고 찾고자 하는 정보는 단순히 검색창에 입력하는 검색어가 아니라 검색 결과에서 실제로 선택하고 이용한 자료를 분석 대상으로 삼으며, 이용자의 관심 주제를 파악할 수 있는 분석방법이다. 예를 들어, 네이버에 입력된 검색어와 클릭로그를 바탕으로 이용자의 검색 행태를 분석한 결과 이용자들은 사이트 검색을 내용 검색보다 많이 실시하였고, 전반적으로 컴퓨터/인터넷, 엔터테인먼트, 쇼핑, 게임, 교육의 순으로 검색을 실시하는 것으로 나타났다.[6]

정리하면, 검색로그 분석은 검색시스템 이용자의 검색 행위에 대한 패턴을 발견하고, 이용자 간의 사회성 관계를 모색하여 추천 서비스를 제공할 수 있다. 이렇듯 검색로그 분석을 기반으로 하는 추천 서비스는 일반적으로 크게 검색어 추천 서비스와 이용자 관계 추천 서비스 두 가지로 구분할 수 있다.

2. 네트워크 분석

네트워크(network)는 다양한 영역에서 사용하는 용어이다. 컴퓨터를 통신회선으로 연결하여 상호 소통하도록 한 것(컴퓨터 네트워크), 사람들이 서로 연결되어 교류하는 것(인맥 네트워크), 정보를 주고받는 것(정보 네트워크), 항공기 노선의 연결구조(항공 네트워크) 등이 네트워크의 주

3) Silverstein, C. et al., "Analysis of a very large web search engine query log," *SIGIR Forum*, Vol.33, No.4 (1999), pp.6-12.
4) 이준호·박소연, "국내 웹 이용자의 검색 행태 추이 분석," 한국문헌정보학회지, 제30권, 제2호(2005), pp.147-160.
5) 이준호·박소연·권혁성, "질의 로그 분석을 통한 네이버 이용자의 검색 형태 연구," 정보관리학회지, 제20권, 제2호(2003), pp.27-42.
6) 박소연·이준호·김지승, "클릭 로그에 근거한 네이버 검색 질의의 형태 및 주제 분석," 한국문헌정보학회지, 제39권, 제1호(2005), pp.265-278.

요한 사례들이다. 네트워크에서는 세상의 모든 것은 연결되어 있다고 하는 연결성 또는 연결주의(connectionism)를 인식론적 기초로 하고 있다. 이러한 연결에는 연결 대상, 연결된 구조, 교류 및 소통의 행위 등이 존재한다.

연결성을 토대로 형성되는 네트워크의 특성은 다음 3가지로 정리가 가능하다. 첫째, 네트워크의 역동성(dynamics)으로 네트워크는 끊임없이 진화하고 자기조직적으로 시스템을 구성한다. 네트워크 안에서는 연결관계가 맺어지기도 하고, 끊어지기도 한다는 것을 의미한다. 둘째, 네트워크의 군집화(clustering)로 네트워크 내의 노드들은 공통의 경험과 관심사에 기초한 작은 군집을 갖는다. 이것은 유유상종의 현상을 설명하는 특징이며, 좁은 세상 네트워크(small-world network)를 형성하는 원인으로 작용한다. 셋째, 네트워크의 중심성(centrality)으로, 네트워크 내에는 영향력이 있는 사람, 정보를 유통시키는 브로커, 기능상의 중심이 되는 자원 등과 같은 소수의 중심집단 또는 핵심세력(허브)이 존재한다는 것이다. 중심성은 네트워크가 중앙(기구)의 지배나 통제가 없이 전체적으로 응집력이 있는 행동이 창발되도록 하는 역할을 한다.

네트워크는 전산학, 생물학, 경제학, 정보학, 사회학 등의 학문 영역에 많이 응용되고 있으며, 실제의 응용 유형에 따라 사회 네트워크, 정보 네트워크, 산업 네트워크(항공 네트워크, 부품 네트워크 등) 등으로 분류가 가능하다. 이 중에서 사회 네트워크(social network)는 사회 연결망으로서, 개인(또는 조직)이 친구 관계, 금전교류 관계, 공동체소속 관계 등과 같은 사회적 관계에 의해 형성된 사회 구조이다. 여기서 노드는 사람(행위자)이며, 링크는 사람과 사람 사이의 사회적 관계이다. 그리고 사회적 관계는 거래 관계, 의사소통 관계, 도구적 관계, 친족 관계 등 다양한 유형이 존재한다.[7)]

네트워크 분석은 사회 네트워크 분석(SNA, Social Network Analysis)으로 알려져 있으며, 분석 수준에 따른 분석방법은 다음의 4가지로 구분이 된다.

첫째, 네트워크 수준 분석으로 거시(macro) 수준의 분석지표를 사용한다. 밀도(density), 포괄성(inclusiveness), 집중도(centralization) 등이 해당되며, 네트워크의 기본적 특성 파악이 가능하다.

둘째, 노드 수준 분석으로 미시(micro) 수준의 분석지표를 사용한다. 인접성 지표와 연결성 지표로 구분되며, 인접성 지표는 연결거리(distance), 직경(diameter) 등, 연결성 지표는 연결정도(degree), 연결강도(strength) 등이 해당된다.

셋째, 네트워크/노드 수준 분석으로 하이브리드(hybrid) 수준의 분석지표를 사용한다. 중심성(centrality) 지표가 대표적이며, 연결정도 중심성(지역 중심성), 근접 중심성(전역 중심성), 매개 중심성이 가장 유명하다. 네트워크 내에서 각 노드들의 영향력 크기를 순위화 하는데 사용이 가능하다.

넷째, 집단(group) 수준 분석으로 노드의 유사성을 기준으로 네트워크의 하위 집단을 분류하고

7) 손동원, 사회 네트워크 분석(서울 : 경문사, 2002), pp.3-4.

해당 집단의 특성을 파악할 수 있다. 집단분석은 군집(clustering) 분석, 구조적 등위성(structural equivalence) 분석, 컴포넌트(component) 분석, 파당(clique) 분석 등으로 구분된다.

가. 네트워크 수준의 기본 속성에 대한 분석

(1) 밀도(density)

한 네트워크에서 노드들 사이에 연결된 정도를 말하며, 전체 노드들이 연결된 개수로 표현된다. 네트워크를 형성하는 관계가 얼마만큼 응집되어 있는지를 설명한다. 밀도가 높은 네트워크는 노드의 연결 관계가 그만큼 많다는 것이고, 서로 긴밀하게 연결되어 상호간 도움이나 교류가 많아지게 된다. 밀도는 0에서 1의 범위 내에서 값을 가진다. 밀도가 0이라는 것은 연결선이 하나도 없는 네트워크이고, 1은 모든 노드들이 연결되어 있다는 의미이다.

(2) 연결거리(distance)

특정 노드의 연결성 단계(또는 차수)의 의미를 가진다. 노드 N1과 N2간의 연결거리가 'd'인 경우, 그들은 d단계 떨어진 거리에 의해 연결되어 있다는 의미이다. N1과 N2간에는 다양한 연결경로가 존재할 수 있다. 따라서 연결거리는 그러한 연결경로 중에서 가장 짧은 거리를 의미한다. 연결거리가 짧을수록 두 노드 간에는 가깝게 연결되어 있으며, 연결성이 높다고 할 수 있다.

(3) 연결정도(degree)

특정한 한 노드의 자체적 속성이며, 해당 노드에 직접 연결되어 있는 노드들의 개수(또는 연결선의 개수)를 말하며, 특정 노드의 영향력 또는 활동력을 인식하는 지표이다. 연결정도가 높으면, 그 노드는 전체 네트워크에서 영향력이 높다는 것이다.

(4) 연결강도(strength)

노드와 노드 사이에 맺어지는 연결의 강도를 말하며, 노드 간 연결의 중요도를 표현한다. 연결강도는 노드 간의 연결에 대한 가치(가중치)로 표현된다. 일반적으로 연결강도의 값이 높으면 '강한 연결 관계(strong ties)', 낮으면 '약한 연결 관계(weak ties)'로 구분한다. 관계의 강도는 두 노드 간의 친밀성(closeness), 지속성, 접촉빈도 등을 의미한다.

나. 중심성 분석

SNA의 분석 지표 중 하나인 중심성(Centrality)은 사회 네트워크에서 개인이 가지는 권력과 영향력이라는 개념과 연결되어 가장 많이 쓰이는 지표이다. 중심성은 "한 행위자가 전체 네트워크에서 중심에 위치하는 정도를 표현하는 지표"로 간단히 정의할 수 있는데, 중심성 분석을 통해 한

네트워크에서 중요한 역할을 하거나 주목받는 행위자가 누구인지, 또 각 행위자들은 그 '중심'에 어느 정도 접근하고 있는지를 알 수 있다.[8][9][10] 중심성의 유형은 관점에 따라 여러 가지로 나눌 수 있는데, 이 중에서 중심성 분석의 가장 기본이 되는 지표는 연결정도, 근접, 매개 중심성이다.

(1) 연결정도 중심성(degree centrality)

네트워크의 노드들이 얼마나 많은 연결을 가지고 있는지를 측정한다. 연결정도 중심성은 한 개인이 더욱 많은 연결을 가질수록 더욱 많은 권력을 가진다는 기본적인 믿음에 근거하여 많은 연결을 가진 개인은 선택의 폭이 넓기 때문에 더욱 많은 기회를 가진다고 보는 것이다. 이러한 자율성은 남에게 덜 의존하게 되고 결국 더욱 강력한 권력을 가질 수 있다고 본다.[11] 네트워크상에서 개인의 권력과 영향력을 단순히 개인이 가진 연결의 수, 개인이 관계하고 있는 사람의 수만으로 평가할 수 있는 것은 아니지만, 해당 개인을 포함하는 지역 수준에서 영향력이 없다고는 할 수 없다.

(2) 근접 중심성(closeness centrality)

한 노드가 얼마나 네트워크의 중앙에 있는지를 측정하여 다양한 노드들이 다른 노드들과의 근접 정도를 보여준다. 근접 중심성을 분석함으로써 네트워크 전역에서 가장 일반적인 영향력을 가지는 노드가 무엇인지를 알 수 있다. 이렇게 네트워크에서 가장 중심이 되는 노드는 자신이 가진 자원을 가장 빠르게 전체 네트워크에 확산시킬 수 있다. 근접 중심성은 한 노드가 다른 노드에 얼마만큼 가깝게 있는가를 보는 것으로 두 노드 사이의 거리를 측정한다. 즉, 근접 중심성이 높은 노드는 네트워크 내 다른 모든 노드와 가장 짧은 경로거리(path length)를 가지고 있어 가장 짧은 시간에 여러 노드에 쉽게 도달할 수 있는 좋은 위치를 차지하고 있다. 이러한 구조적 강점은 해당 노드에게 권력을 가져다 줄 수 있다고 본다.

(3) 매개 중심성(betweenness centrality)

한 노드가 다른 노드와 네트워크를 구축하는데 있어 중개자 혹은 다리 역할을 얼마나 수행하느냐를 측정하는 개념으로 중개 역할을 '중심'으로 간주할 때 사용한다.[12] 매개 중심성은 한 노드가 네트워크 내의 다른 노드들 사이에 위치하는 정도를 측정하는 것으로, 이 위치에 있는 사람 혹은 기관은 정보의 흐름에 있어 큰 영향력을 가질 수 있다. 이러한 노드의 존재는 네트워크가 전반적으

8) 앞의 책, p.93.
9) 이재윤, "계량서지적 네트워크 분석을 위한 중심성 척도에 관한 연구," 한국문헌정보학회지, 제40권, 3호(2006), pp.191-214.
10) Drew Mackie, "Network mapping," 〈http://www.williemiller.co.uk/network-mapping.htm〉 [cited 2009. 9. 20].
11) Robert A. Hanneman, "10. Centrality and power," at Introduction to social network methods, 〈http://faculty.ucr.edu/~hanneman/nettext/Introduction_to_Social_Network_Methods.pdf〉 [cited 2009. 5. 27].
12) 손동원, 앞의 책, p.95.

로 잘 연결되어 있다는 것을 보증하기도 하지만 그들이 잠재적으로 그들 자신의 의제(agenda)에 따라 정보를 필터링(filtering)하거나 네트워크를 취약하게 만들 수 있는 위험을 내포하고 있다.

다. 군집 분석

군집 분석은 유사한 특성을 가지는 데이터들을 함께 모아서 이들 데이터가 가지고 있는 공통적인 특성으로 전체의 개체들을 몇 개의 그룹 또는 군집 또는 클러스터(cluster)로 나누는 것을 말한다. 이때 군집 또는 클러스터란 유사성(similarity)을 지니는 개체들로 구성된 일련의 데이터들을 뜻한다. 즉 군집 분석은 흩어져 있는 여러 데이터들에 대한 군집화의 과정이며, 정보학 및 데이터마이닝 분야에서는 '클러스터링' 또는 '클러스터 분석'으로 불리며, 통계학을 비롯한 다양한 분야에서는 '집단 분석'으로 주로 표현된다.

군집 분석은 대용량 데이터, 그 자체에만 의존하여 자료 탐색과 자료를 요약하는, 사전에 정의된 어떠한 특수한 목적이 없는 자료 분석 기법이다. 즉, 전체 데이터를 군집을 통해 잘 구분하는 것이 분석의 최대 목적이라 할 수 있다. 따라서 동일한 군집의 개체들은 유사한 성격을 갖도록, 서로 다른 군집에 속한 개체들 사이에는 그와 반대로 상대적으로 다른 성격을 갖도록 군집이 형성되어야 할 것이다.

군집의 기법은 계층적 기법과 비계층적 기법으로 구분된다. 계층적 기법으로는 단일 연결(single linkage), 완전 연결(complete linkage), 그룹 평균 연결(group average linkage), 워드 기법(Word's method) 등이 있다. 비계층적 기법으로는 싱글패스(single pass), K-means, EM(expectation maximization) 알고리즘 등이 있다.[13] 계층적 기법이 비계층적 기법에 비해 클러스터링이 잘되며, 비계층적 기법은 계층적 기법에 비해 군집 처리 시간이 짧다고 알려져 있다. 일반적으로 클러스터링 성능을 위하여 계층적 기법을 선호하지만, 대규모 문서처리를 위해서 비계층적 기법의 이용이 효율적이다. 계층적 기법 중의 하나인 단일연결기법은 대형 군집을 생성하기 때문에 문서 클러스터링에 적합하지 않다고 볼 수 있다.

라. 에고 네트워크 분석

네트워크의 분석은 전체 네트워크를 대상으로 하는 분석이 아니라, 특정한 노드를 중심으로 하는 에고 네트워크 분석(ego network analysis)이 필요한 경우가 있다. 여기서 에고(ego)는 개인, 집단, 조직, 사회 등을 의미하며, 에고 네트워크 분석은 에고(focal node)와 타자(alter)들 간의 '사회적 관계'에 의한 연결 상황을 분석하는 것을 말한다.

에고 네트워크의 주요 개념 및 분석 지표는 에고의 적정한 경로길이에 있는 타자들의 집합인

13) 정영미·이재윤, "지식 분류의 자동화를 위한 클러스터링 모형 연구," 정보관리학회지, 제18권, 제2호(2001), pp.203-230.

이웃노드(neighborhood), 경로길이 N인 이웃 타자들(N-step neighborhood), 크기, 연결정도, 직경 등과 같은 밀도(density), 브로커(broker), 에고의 매개 중심성(ego betweenness), 중복이 적은 노드를 의미하는 구조적 틈새(structural hole) 분석 등이 있다.

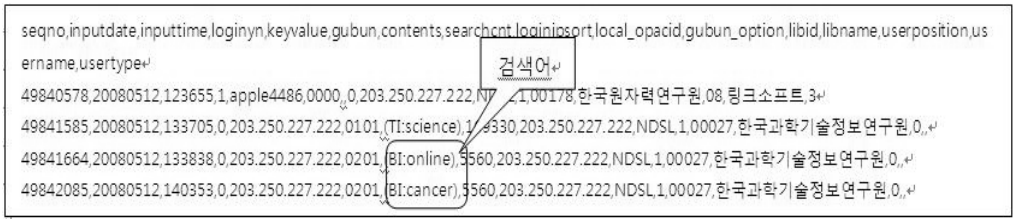
Ⅲ. 실험의 설계와 데이터 처리

검색로그 네트워크의 분석을 위한 실험은 2008년 10월부터 12월까지 총 3개월 간 NDSL에 로그인하여 검색한 이용자들과 검색어들을 대상으로 하였다. 이 기간 동안 전체 이용자는 14,658명이었으며, 데이터의 처리와 분석을 원활하게 하기 위하여, 검색 세션이 50건 이상인 351명과 이들이 사용한 검색어 17,790건으로 실제의 실험 데이터를 제한하였다.

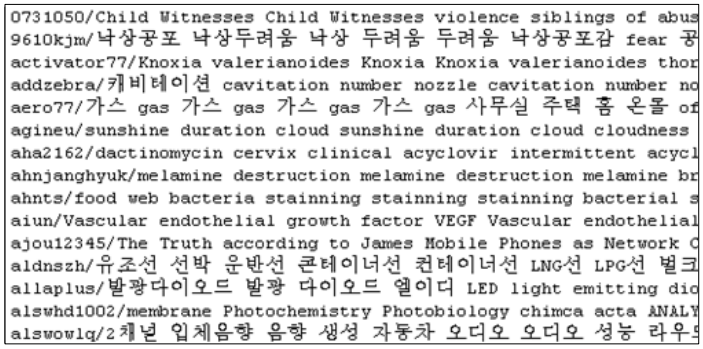
실험의 처리 과정은 전체적으로 전처리 과정, 네트워크 생성 과정, 네트워크 분석 과정으로 나누어진다. 전처리하는 과정에서 REGEXBUDDY 3.1.0, EXCEL 2007, MYSQL 5.0 등의 데이터처리 소프트웨어를 사용하였고, 네트워크의 분석과 시각화를 위한 도구로는 UCINET 6과 PAJEK 1.25를 선택하였다.

1. 전처리 과정

전처리 과정에서는 검색로그 데이터를 입력받아 연관행렬을 출력하는 작업을 수행한다. 세부적인 처리작업은 검색로그 추출/정제 작업, 형태소분석 작업, 연관행렬 작성 작업으로 구분된다. NDSL 검색서버로부터 얻은 검색로그 데이터는 〈그림 1〉과 같다. 이 데이터에서 검색자와 검색어를 추출하여 리스트를 만든다. 그런 다음 형태소분석기를 통하여 입력된 검색어를 구성하는 각각의 형태소 중 불규칙 활용이나 축약, 탈락 현상이 일어난 경우에는 원형을 복원하는 과정을 거친다. 그리고 검색어에 대하여 의미 없는 단어, 즉 불용어를 제거하고 진행하였다. 이러한 작업에서 얻어진 검색어 리스트는 〈그림 2〉와 같다.



〈그림 1〉 NDSL 검색로그 데이터 예시



〈그림 2〉 전처리 단계에서 추출한 검색어 리스트

로그 데이터에서 나온 검색어 리스트에서 검색어의 출현빈도를 바탕으로 〈그림 3〉과 같은 검색자 × 검색어 가중치 행렬을 작성한다. 가중치 계산은 정보검색분야에서 많이 쓰이는 TF*IDF 기법을 사용하였다.

1		Child	Witnesses	violence	siblings	abuse	Famaily	Vic
2	731050	2	2	36	1	1	1	
3	9610kjm	0	0	0	0	0	0	
4	activator7	0	0	0	0	0	0	
5	addzebra	0	0	0	0	0	0	
6	aero77	0	0	0	0	0	0	
7	agineu	0	0	0	0	0	0	
8	aha2162	0	0	0	0	0	0	
9	ahnjanghy	0	0	0	0	0	0	
10	ahnts	0	0	0	0	0	0	
11	aiun	0	0	0	0	0	0	
12	ajou12345	0	0	0	0	0	0	
13	aldnszh	0	0	0	0	0	0	
14	allaplus	0	0	0	0	0	0	

〈그림 3〉 검색자×검색어의 이원모드 행렬(일부)

2. 네트워크 생성 과정

검색자×검색어의 이원모드 행렬에서, 가장 많이 알려진 유사계상 공식인 같은 코사인(cosine) 유사계수를 이용한 이용자 간 유사도를 계산하여 〈그림 4〉와 같은 검색자×검색자 간 일원모드 행렬로 변환한다.

1		0731050	9610kjm	activator7	addzebra	aero77
2	0731050	2741	0	0	0	0
3	9610kjm	0	585	0	0	0
4	activator77	0	0	223	15	0
5	addzebra	0	0	15	1426	10
6	aero77	0	0	0	10	882
7	agineu	0	0	0	0	0
8	aha2162	0	0	0	3	0
9	ahnjanghyuk	0	0	8	16	0
10	ahnts	0	0	0	4	15
11	aiun	0	0	0	0	0
12	ajou12345	0	0	0	0	0

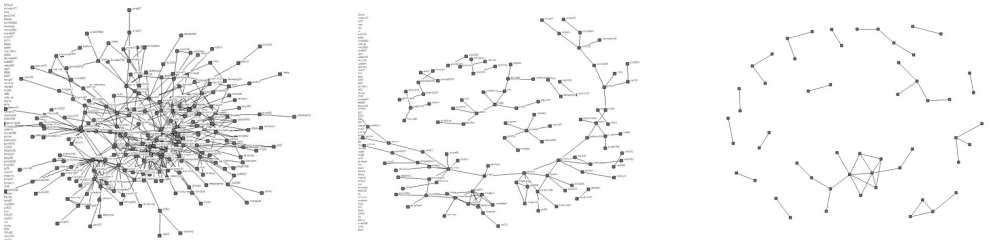
〈그림 4〉 검색자×검색자 일원모드 행렬(일부)

코사인 유사계수의 값은 0에서부터 600 이상까지 아주 광범위한 분포를 나타내고 있다. 따라서 네트워크의 연결성을 판단하는 기준 임계치(critical point)를 60 이상에서부터 600 이상까지 단계적으로 조절하여 네트워크를 생성하고, 해당 네트워크의 구조 변화를 관찰하였다. 즉, 전체 링크노드의 수가 급격하게 변화하고, 구조가 선명하게 드러나는 시점은 〈표 1〉에서와 같이 임계치가 100에서 300 이상인 경우임을 알 수 있었다.

〈표 1〉 임계치 기준에 따른 네트워크의 기본적 속성

유형	임계치 기준	링크 노드수	전체 링크수	평균 링크수	밀도
1	20 이상	344	8212	23.87	0.070
2	40 이상	319	3720	11.66	0.036
3	60 이상	291	2030	6.98	0.024
4	80 이상	254	1314	5.17	0.020
5	100 이상	223	880	3.95	0.017
6	200 이상	114	234	2.05	0.018
7	300 이상	54	88	1.63	0.031
8	400 이상	37	54	1.46	0.041
9	500 이상	24	38	1.58	0.069
10	600 이상	17	26	1.53	0.096

〈그림 5〉는 3개월 간 NDSL에서 논문을 검색한 이용자들 중 검색 세션이 50건 이상인 351명의 네트워크이며, 임계치가 100 이상, 200 이상, 300 이상인 경우로 구분한 연결 관계를 나타내고 있다. 이 중에서 임계치가 200 이상인 경우와 100 이상인 경우의 검색자 네트워크가 보다 선명하게 나타남을 알 수 있다.



a) 임계치가 100 이상 b) 임계치가 200 이상 c) 임계치가 300 이상

〈그림 5〉 검색자 네트워크의 구조

Ⅳ. 네트워크 분석

1. 중심성 분석

앞서 생성한 검색자 네트워크에서 연결정도 중심성, 근접 중심성, 매개 중심성으로 나누어진 검색자의 중심성을 분석하였다. 먼저, 3가지 유형의 임계치 기준에 따른 중심성 값이 높은 상위 10명의 검색자 리스트를 도출한 결과는 〈표 2〉에서 〈표 4〉에 정리하였다.

검색자 네트워크에서 검색자의 연결정도 중심성은 인접의 검색자들과 공유하는 검색어가 많아서, 지역적 수준에서 허브 역할을 하는 검색자를 의미한다. 이 중에서 로그인 ID가 sukmp, sold77, tima21인 검색자들이 높은 연결정도 중심성 값을 나타내었다. 근접 중심성 값이 상위인 검색자는 sukmp, viper, tima21 등이며, 이들은 검색자 네트워크의 중앙에 위치하여 네트워크 내의 다른 검색자들과 근접하며, 전역적 수준에서 허브 역할을 하고 있다. 매개 중심성 값이 높은 검색자도 sukmp, viper, tima21 등이며, 이들은 검색어를 통해 다른 검색자들과 중개해 주는 역할이 많이 한다고 볼 수 있다.

〈표 2〉 검색자 중심성 순위(상위 10명) - 임계치 기준 100

순위	연결정도 중심성		근접 중심성		매개 중심성	
	검색자	값	검색자	값	검색자	값
1	sukmp	0.102679	sukmp	0.366101	sukmp	0.241976
2	ayusuj	0.09375	viper	0.349804	mundo00	0.159265
3	mundo00	0.089286	ayusuj	0.344869	ayusuj	0.132838
4	viper	0.075893	smyang38	0.334896	smyang38	0.129086
5	lilis	0.071429	hwp73	0.323569	viper	0.097172
6	zi882	0.071429	mirrone	0.315225	zi882	0.081128
7	sold77	0.0625	zi882	0.313875	sold77	0.07961
8	gbg999	0.058036	lilis	0.312983	yeje3749	0.076473
9	mirrone	0.058036	sold77	0.311653	yyook	0.069834
10	yys18	0.058036	hjkasu	0.311212	yys18	0.068325

〈표 3〉 검색자 중심성 순위(상위 10명) - 임계치 기준 200

순위	연결정도 중심성		근접 중심성		매개 중심성	
	검색자	값	검색자	값	검색자	값
1	sold77	0.034188	viper	0.178466	viper	0.158626
2	sukmp	0.034188	ayusuj	0.1683	ayusuj	0.156453
3	tima21	0.029915	snikerskr	0.141119	sukmp	0.14457
4	ayusuj	0.025641	sukmp	0.140697	snikerskr	0.14336
5	snikerskr	0.025641	guevara13	0.10256	hjkasu	0.135428
6	chajh73	0.021368	xi010	0.088654	sold77	0.135428
7	hjkasu	0.021368	yyook	0.085809	saintsky80	0.125506
8	mundo00	0.021368	sold77	0.079177	guevara13	0.125048
9	omaeya	0.021368	jisim1004	0.053097	gbg999	0.123694
10	skybaek	0.021368	hw0726	0.046302	genman	0.123694

〈표 4〉 검색자 중심성 순위(상위 10명) - 임계치 기준 300

순위	연결정도 중심성		근접 중심성		매개 중심성	
	검색자	값	검색자	값	검색자	값
1	sold77	0.068182	tima21	0.144444	tima21	0.0292695
2	tima21	0.068182	sold77	0.131313	mundo00	0.021045
3	chajh73	0.034091	mundo00	0.115556	sold77	0.0176584
4	chemlove	0.034091	chajh73	0.111111	smyang38	0.0079826
5	hjkasu	0.034091	omaeya	0.111111	viper	0.0079826
6	mapleleaf	0.034091	aldnszh	0.106996	zi882	0.0036284
7	mundo00	0.034091	mapleleaf	0.09319	hjkasu	0.0021771
8	omaeya	0.034091	murima	0.09319	pelvic	0.0021771
9	zi882	0.034091	viper	0.09319	chajh73	0.0014514
10	aldnszh	0.022727	smyang38	0.084967	chemlove	0.0014514

위 중심성 분석의 결과, sukmp, sold77, viper, tima21 등은 중심성 값이 상대적으로 높은 중심 검색자임을 알 수 있다. 따라서 이들이 사용한 검색어들은 무엇이며, 어떤 검색어들에 의해 연결되어 있으며, 중심성 값이 높게 나오도록 작용한 원인은 무엇인지 파악해볼 필요가 있다. sukmp는 전체 84건의 검색어를 사용하였으며, 대부분의 검색어가 영어로 되어 있었다. sold77는 전체 425건의 검색어를 사용하였으며, 이 중에서 한글 검색어는 165건이었다. viper는 전체 156건의 검색어를 사용하였으며, 대부분이 영어의 검색어였다. 그리고 tima21은 전체 311건의 검색어를 사용하였으며, 이중에서 한글 검색어는 170건이었다. 이들이 사용한 검색어들의 리스트 일부는 〈표 5〉에 정리되어 있다.

중심성이 높다는 것은 검색자들이 공유하는 검색어들이 많다는 것을 의미하며, 이는 관심을 가지는

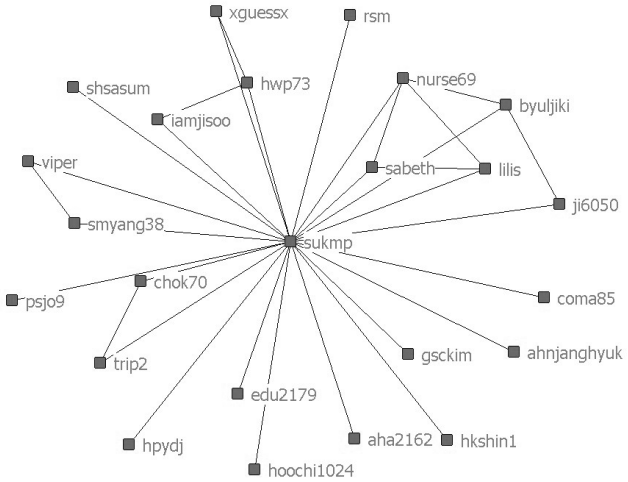
주제가 유사하다고 판단할 수 있다. 표에서 알 수 있듯이, sukmp와 viper가 공유하는 검색어가 있고, sold77과 viper도 검색어를 공유하고 있다. 자세히 살펴보면, sukmp는 viper 사이에 'chamber', 'measurement', 'system' 등 전체 5건의 검색어, sold77과 viper는 'carbon', 'energy', 'gas', 'plant' 등 전체 7건의 검색어, 그리고 sold77과 tima21과 '선박', '디젤', '추진축', 'engine', 'propulsion' 등 46건의 검색어와 공유하는 것으로 나타났다.

〈표 5〉 중심성 상위 검색자 4명의 검색어 리스트 일부(괄호는 출현 빈도수)

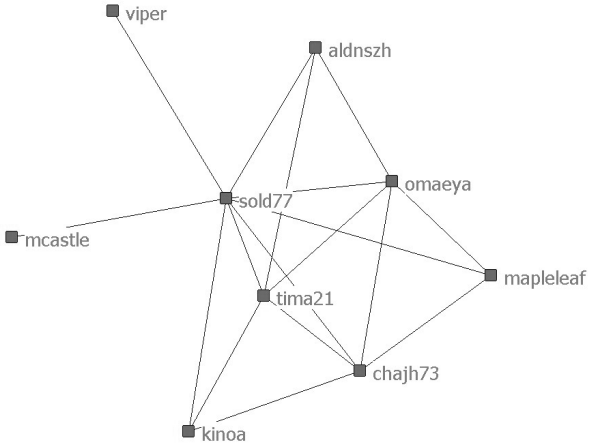
검색자	검색어 리스트(빈도수 〉5)
sukmp (84)	radiation(48), gel(36), optical(31), therapy(25), dosimetry(18), CT(11), dimensiona(6), ct(6), imaging(5), experimental(5), ultrasound(5), tomography(5), chamber(5), dimension(5), phantom(5), igrt(5), radiotherapy(5), measurement(4), system(2)
sold77 (425)	saving(29), efficiency(28), engine(22), fuel(20), propulsion(20), plant(16), hull(16), 선박(12), energy(11), form(11), ship(10), 에너지(10), Energy(10), 추진(10), Vessel(9), 설계(9), 디젤(9), Engine(7), consumption(7), 후처리장치(7), 추진축(7), design(6), 날개(6), FUEL(6), 해석(6), 프로펠러(6), 유동(6), carbon(1), gas(2)
viper (156)	exchange(26), chamber(23), canopy(22), gas(20), carbon(16), plant(15), system(14), field(14), measurement(12), CO(10), crop(10), set(10), respiration(10), mulching(10), photosynthetic(9), leaf(8), automated(8), whole(8), open(7), dioxide(7), net(7), evapotranspiration(6), soil(6), experimental(6), closed(6), rates(6), soybean(6), global(5), response(5), measuring(5), portable(5), photosynthesis(5), energy(1),
tima21 (311)	계(22), 선박(21), 추진축(17), 진동(13), 해석(12), 비(11), ship(10), 디젤(10), 베어링(9), 틀림(9), 특성(8), shaft(7), torsional(7), 축(6), systems(6), 종(6), 설계(6), vibrations(5), Torsional(5), 탄성(5), Shafting(5), propulsion(2), engine(1)

2. 에고 네트워크 분석

중심성 값이 높은 sukmp, sold77, viper, tima21의 4명의 검색자들의 연결구조는 에고 네트워크를 통해서도 살펴볼 수 있다. 〈그림 6〉과 〈그림 7〉은 sukmp(임계치 100 이상)와 sold77(임계치 200 이상)의 에고 네트워크이다. 즉, sukmp는 viper 등 23명의 이웃노드들과 연결되어 있고, sold77는 viper와 tima21 등 8명의 이웃노드들과 연결되어 있다. 특히 sold77과 tima21은 앞서 살펴본 대로 공유 키워드가 매우 많으므로, 에고 네트워크에서도 가깝게 연결되어 있음을 알 수 있다.



〈그림 6〉 sukmp의 에고 네트워크(임계치 100 이상)



〈그림 7〉 sold77의 에고 네트워크(임계치 200 이상)

3. 군집 분석

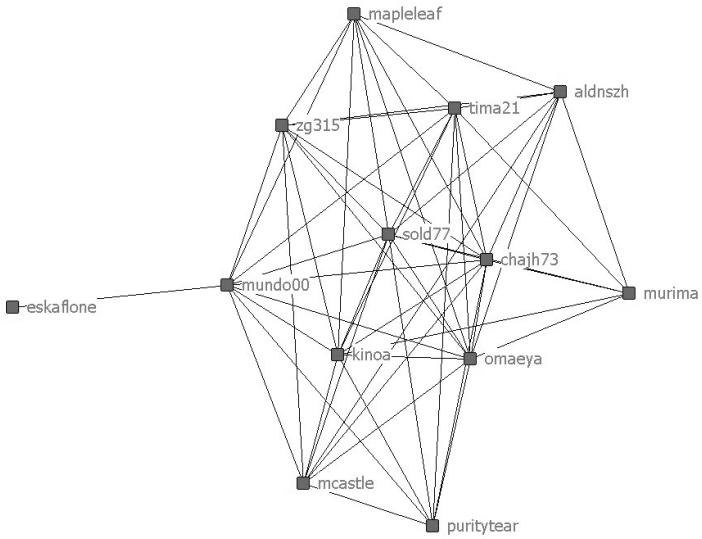
군집 분석은 검색로그 네트워크에서의 검색자들이 어떻게 하위집단으로 구분되어 있는지를 확인하는 데 유용하다. 그리고 특정한 집단을 선정하여 그 집단 내에서 검색자와 검색어가 형성하는 네트워크의 구조를 확인하여 검색자나 검색어를 추천하는 데 활용도 가능하다. 따라서 전체 검색자 네트워크(임계치 200 이상)를 대상으로, 계층적 군집기법인 워드기법(Wards hierarchical clustering)을 이용하여 군집분석을 수행하였다. 집단의 개수는 Dixon-Kronmal 공식을 이용하였으며, 그 결과

〈표 6〉과 같이 21개의 집단이 도출되었다. 집단 내 검색자는 2명에서부터 13명까지 다양하게 분포하고 있다.

〈표 6〉 검색자 네트워크에서의 군집분석 결과

집단	검색자수	검색자 리스트
1	10	0731050, ayusuj, dzero77, funkyrose, gbg999, hjkasu, kingpin22, psjo9, saintsky80, vagosang
2	12	addzebra, byuljiki, genman, hjs2122, hkshin1, hwp73, kcbay, nurse69, sabeth, smyang38, sukmp, viper
3	8	aha2162, bach36, chemlove, chemy9, guevara13, jordam, mlee9, snikerskr
4	4	ahnjanghyuk, bahk0527, icbk8, ymiran
5	13	aldnszh, chajh73, eskaflone, kinoa, mapleleaf, mcastle, mundo00, murima, omaeya, puritytear, sold77, tima21, zg315
6	13	atroforce, ckdpharm, coelacanth19, cup, edu2179, gusdud, k3t62, leejees, lilis, mirrone, pelvic, star0717, zi882
7	2	avaco, saeib
8	2	chok70, ike2000
9	4	coolpig0416, gomting48, green2525, munkyo
10	7	dooly9842, omgboy, rezzo79, savant, sdh791, skybaek, vocaleo
11	4	forsky80, iattain, jereintchi, yeje3749
12	2	geobest, hardiron77
13	4	gkgkmouse, jungdk706, kaist, trip2
14	2	hey4194, kaykim25
15	4	hurtur, jaiwook, parkjw, xguessx
16	12	hw0726, jisim1004, jisooyu, jiwj0503, ksltv, mine27, namgreen18, okmadam, xi010, yshau75, yyook, yys18
17	2	jyoungmee, phs0585
18	2	jyounjo0707, pulyc
19	2	katrino, shyunmd
20	3	kjh2336, pp124pp, yjyhappy
21	2	leehaeng19, libkj

다음은 21개 집단 중에서 5번째 집단(집단-5)을 선정하여, 주제적 유사성을 가지는 검색자 집단이 나타내는 구체적인 특성을 파악하였다. 이 집단은 소속 검색자 수가 가장 많으며, 앞서 분석한 검색자 네트워크의 중심성 값이 높은 검색자인 sold77, tima21가 포함되어 있기에 참조적인 분석이 가능하다고 판단하였다. 이들을 포함한 전체 13명의 검색자가 사용한 검색어는 전체 2,236건이다. 전체 링크의 수는 123개이며, 네트워크의 밀도는 0.728로 나타났다. 〈그림 8〉은 집단-5의 검색자 네트워크의 구조이며, 〈표 7〉은 이들 13명 검색자들에 대한 중심성을 분석한 결과이다. 즉, 집단-5에서는 동일 집단 내 검색자들이기에 상호간에는 아주 밀접하게 연결되어 있지만, 이 중에서 chajh73, mundo00, omaeya, sold77 등이 중심 검색자임을 알 수 있다.



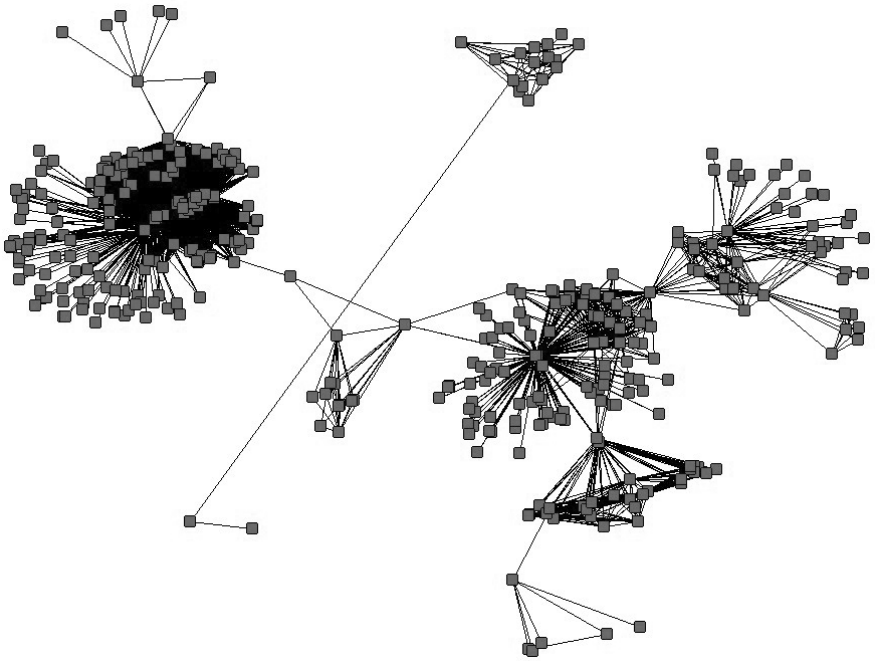
〈그림 8〉 집단-5에서의 검색자 네트워크

〈표 7〉 집단-5 검색자들의 중심성 순위(상위 10개)

순위	연결정도 중심성		근접 중심성		매개 중심성	
	검색자	값	검색자	값	검색자	값
1	chajh73	0.916667	chajh73	0.923077	mundo00	0.175415
2	omaeya	0.916667	omaeya	0.923077	chajh73	0.035173
3	sold77	0.916667	sold77	0.923077	omaeya	0.035173
4	kinoa	0.833333	kinoa	0.857143	sold77	0.035173
5	mundo00	0.833333	mundo00	0.857143	tima21	0.028084
6	tima21	0.833333	tima21	0.857143	kinoa	0.025920
7	zg315	0.750000	zg315	0.800000	aldnszh	0.012139
8	aldnszh	0.666667	mapleleaf	0.750000	mcastle	0.011418
9	mapleleaf	0.666667	mcastle	0.750000	zg315	0.010281
10	mcastle	0.666667	aldnszh	0.705882	mapleleaf	0.006223

가. 검색어 네트워크의 분석

집단-5에서 검색어와 검색어 간의 연관도를 일정한 계산방법(TF*IDF, 코사인 유사계수 등)을 이용하여 측정하여 검색어 네트워크를 도출하였다. 전체 930개 검색어 중에서 네트워크의 구조가 선명하게 나타나는 시점은 임계치가 100 이상인 경우이며, 〈그림 9〉와 같은 네트워크의 구조(노드 수=383개, 링크 수=7,040, 밀도=0.049)가 나타났다. 〈표 8〉은 이 네트워크의 중심성을 분석한 결과이다. '할로겐', 'efficiency', 'BDE', 'saving', '설계' 등의 검색어가 집단-5를 구성하는 검색자 네트워크의 중심 검색어임을 알 수 있다.



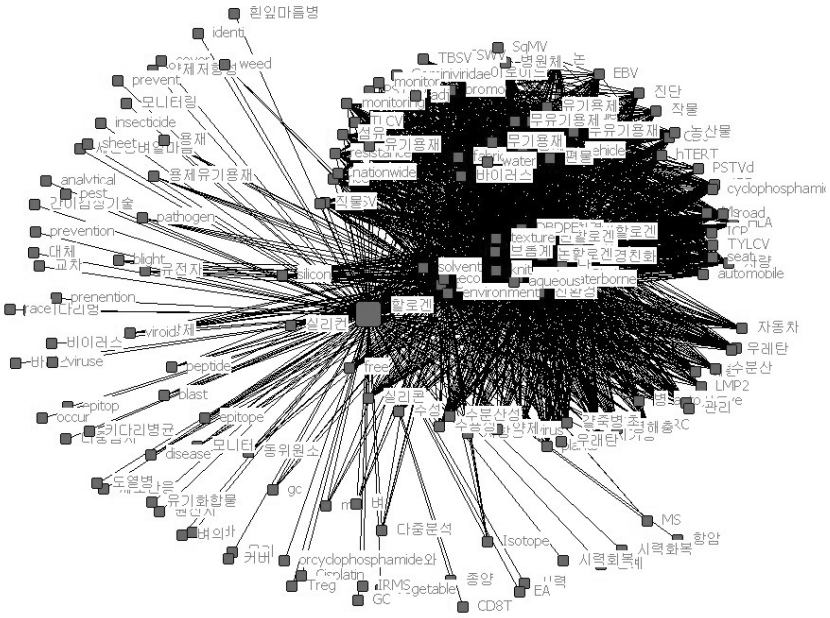
〈그림 9〉 집단-5의 검색어 네트워크의 구조(임계치-100 이상)

〈표 8〉 집단-5의 검색어 네트워크의 중심성 순위(상위 5개)

순위	연결정도 중심성		근접 중심성		매개 중심성	
	검색어	값	검색어	값	검색어	값
1	할로겐	0.434555	efficiency	0.33905	할로겐	0.55736
2	BDE	0.308901	saving	0.336093	설계	0.454068
3	environment	0.308901	설계	0.33415	efficiency	0.28553
4	eco	0.303665	할로겐	0.333829	saving	0.255807
5	solvent	0.303665	3차원	0.300562	plant	0.208875

나. 검색어의 에고 네트워크 분석

집단-5에서 중심성 값이 가장 상위에 있는 '할로겐'이라는 검색어가 나타내는 에고 네트워크는 〈그림 10〉과 같다. 이 검색어는 하위집단 내에 있는 대부분의 검색어들과 직접 연결되어 있음을 알 수 있다.



〈그림 10〉 '할로겐' 검색어의 에고 네트워크

V. 결 론

지금까지 검색로그에 나타난 검색자들의 네트워크 연결구조를 파악하고, 네트워크 분석방법을 통해 검색자들의 검색행위에 드러난 다양한 특성들을 살펴보았다. 이러한 분석작업을 통해 얻어진 결과를 정리하면 다음과 같다. 첫째, 검색자들은 검색어의 유사성에 따라 네트워크라는 연결구조를 나타냄을 알 수 있었다. 연결관계는 공유하는 검색어에 따라 유사성을 계산하여 판단하였으며, 유사성을 판단하는 기준(임계치)에 따라 다양한 연결구조의 네트워크가 생성되었다. 둘째, 특정한 유사성 기준에 의한 검색자 네트워크에서 중심성을 분석할 결과, 연결정도 중심성, 근접 중심성, 매개중심성에 의한 중심 검색자(허브)가 존재함을 알 수 있었다. 이들은 내트워크에서 지역적, 전역적, 매개적 관점에서 존재하는 중심이며, 이들이 사용한 검색어는 향후 연관검색어 추천기법을 개발하는데 유용하게 사용할 수 있을 것이다. 셋째, 중심 검색자들에 대한 에고 네트워크는 그들의 이웃 검색자들이 누구인지 파악이 가능하며, 그들이 공유하는 키워드의 존재를 파악할 수 있었다. 넷째, 전체 검색자들이 사용한 검색어의 유사성에 따라 군집분석을 시도한 결과, 검색자 네트워크는 다수의 집단이 존재함을 확인하였다. 하나의 집단 내에 속하는 검색자들은 공유하는 검색어가 많으므로, 밀도가 높은 네트워크로 표현되었다. 그리고 그들이 사용한 검색어들도 연결정도가 상당히 높

은 네트워크로 나타남을 알 수 있었다.

본 연구의 실험은 NDSL의 3개월간 검색로그 데이터에서 검색빈도가 많은 일부의 검색자들을 대상으로 하였다. 보다 장기간에 나타난 검색로그 데이터와 전체 검색자들을 분석대상으로 할 경우, 본 연구와는 차이가 있는 결과를 나타날 수도 있을 것이다. 그러나 본 연구는 검색로그 데이터에 나타난 검색자들의 관심 주제에 따른 사회적 연결구조를 파악하고자 하였으며, 관심 검색어에 따라 검색자들이 네트워크로 연결될 수 있음을 확인하였기에 향후의 연구로 나아가는 발판은 어느 정도 마련하였다고 본다. 따라서 후속연구는 연구의 기간과 대상을 확대하고, 검색자들의 네트워크에 나타나는 다양한 속성들을 규명하고, 이를 활용한 연관된 검색자와 검색어를 추천하는 보다 정교한 알고리즘을 개발하는 것으로 발전하여야 할 것이다.

〈참고문헌은 각주로 대신함〉