Journal of the Research Institute of Industrial Technology, Vol. 18, 2001

국내 Web Search Agent 평가에 관한 연구

신창훈[1], 이지훈[2], 류병무[2]

A Study of the Performance Evaluation for World Wide Web Search Agents of Korea

C. H. Shin, J. H. Lee, B. M. Ryu

Abstract

This research focuses on the performance evaluation for world wide web search agents of Korea. As the search engines normally serve us with web-directory and web-index methods, we evaluate the performance of the search agents in each method. First, we evaluate the web-directory method with the correlation between keyword and directory. Second, we evaluate the web-index method and examine the complementary relationship among the search agents. As a result, we present an effective method for searching information. We show such a way that a search site can be operated more effectively.

1. 서 론

제 1절 연구의 목적

사용자가 원하는 정보를 쉽게 찾아주는 검색 서비스는 1994년 야후(Yahoo!)가 처음으로 등장한 이래 지금까지 국내외적으로 급격히 증가하고 있다. 이렇게 검색 서비스를 제공하는 사이트가 증가할수록 인터넷 검색에 있어서 전문적인 지식이나 감각을 사용자로 하여금 요구하기도 한다. 또한, '어떤 검색 서비스를 사용해야 할 것인가' 라는 선택의 문제도 발생한다. 실제로 정보검색 방법이나 정보검색엔진의 분류, 각 검색엔진의 특징 등을 파악하는 것은 인터넷 정보검색에 있어 중요하다.

1) 한국해양대학교 물류시스템공학과
2) 한국해양대학교 대학원

대부분의 일반인들이 사용하는 검색 서비스들은 구체적이고 객관적인 검색엔진의 성능에 비하여 사용측면에 있어서는 쉬운 검색 서비스를 선호한다. 이러한 문제를 해결하기 위해서 국외의 경우 이용자가 원하는 키워드(Key-Word)에 따라 가장 적절한 검색결과를 보여주기 위해 많은 연구가 이루어져 왔으며, 단순히 이용자 설문 등을 통하여 검색 엔진들의 성능을 비교, 평가 및 개선하는 데는 한계가 있음을 인지하고 있다.

본 연구의 목적은 검색엔진의 중요 부분인 검색 에이전트(Agent)를 보다 객관적인 데이터를 사용해서 평가하고 좀더 효율적인 검색방법을 모색하고자 하는데 있다. 이를 통해 궁극적으로 이용자에게 보다 정확하고 효율적인 정보를 전달할 수 있는 방법을 모색하는데 기여하고자 한다.

제 2 절 연구 방법

본 연구에서는 각 검색 에이전트를 Web-Directory 방식과 Web-Index 방식으로 분류하여 다음과 같은 방법으로 유사성을 파악 및 평가한다. 첫째, Web-Directory 방식에서는 각 검색 에이전트의 키워드 검색결과에서 분류 되어지는 디렉토리를 이용하여 키워드와 디렉토리간의 관련성을 살펴본다. 둘째, Web-Index 방식에서는 각 검색 에이전트의 키워드 검색을 통한 결과에서 중복되어지는 URL 의 빈도를 측정하여 각 검색 에이전트의 유사성 정도를 측정 및 평가한다.

연구목적 달성을 위하여 우선 기존의 연구를 바탕으로 다음과 같은 단계로 분석을 한다. 우선 첫 번째로 검색엔진에 입력할 키워드를 선정 하고, 두 번째로 실험 검색 사이트를 선정한다. 마지막으로 선택되어진 검색 사이트에 키워드 검색 결과의 URL(Uniform Resource Locator) 데이터를 분석하여 검색 에이전트를 비교평가 한다.

제 3 절 선행 연구

1.1 검색 에이전트의 선행 연구

인공지능 분야에서 관심을 가지고 연구를 시작한 에이전트의 개념을 웹 환경에서 고객대서버(Customer Vs. Server; CS)의 대화 도구로 이용하고자 하는 연구가 90 년대 중반에 진행되었다. 지금까지 에이전트란 대략적으로 사용자를 대신하여 유용한 작업의 일부를 실행하는 도구의 의미를 가지고 있다.

최초의 웹 에이전트는 1993 년 Mattew Gray 가 전 세계의 웹 서버가 몇 개일까라는 궁금증으로 만든 ' World Wide Wanderer' 이며 이후 웹 에이전트에 대한 많은 연구가 진행되고 있다. 웹 에이전트의 기능은 첫째, 에이전트는 대화와 협력으로 다른 에이전트에게 도움을 청하거나 몇몇 에이전트와 한데 어울려 어떤 문제를 풀기 위해 협력할 수 있다. 둘째로 이동기능을 지니고 있어 웹 공간의 여러 기종의 호스트들간을 이동하며, 세 번째 기능으로 추론과 판단기능을 지니고 있어 환경을 인식하고 적절히 판단하여 스스로 어떤 행동을 취하는 능동성을 가지며 학습을 통해서 환경에 적응하는 기능을 가지고 있다.

미국을 비롯한 선진국에서는 에이전트에 기반 한 정보검색 시스템이 각 특성과 해당 서비스의 종류에 따라 Recommendation Agent, New-contents Agent, Search Agent, Customized Agent, Personal-status Agent 등 여러 가지로 개발되고 있다.[1] 최근에는 에이전트의 이론과 구조, 언어, 응용 등으로 많은 연구가 진행 중에 있다.

1.2 검색 에이전트 평가의 선행 연구

검색 에이전트에 대한 평가는 1996 년 이후 꾸준히 계속되어 오고 있으며 각기 다양한 키워드를 사용하여 연구가 진행되어 왔다. Westera(1996)는 단지 ' Wine' 에 관한 검색어를 통해 검색엔진의 성능평가에 관한 연구를 하였고, Feldman(1997)은 ' Car, Tennis Elbow, information retrieval' 등의 다양한 질의를 통해서 연구를 하였다.

이와 같이 대부분의 연구들은 검색어를 통해서 수집된 자료를 바탕으로 진행되어 왔다. 그리고 이러한 검색어의 선택은 관련 서적, 전문가의 의견 및 통계자료 등을 바탕으로 선택하였다. 이런 한정된 키워드로 검색엔진을 전반적으로 평가한다는 데에는 한계가 있으나, 검색어를 통해서 검색엔진의 대략적인 특성을 파악하는 데는 효과적이라 할 수 있다. 이렇게 다양한 검색엔진에 동일한 키워드를 입력하여 나온 데이터를 분석하는 연구는 꾸준히 진행되어 왔다.

Lake, M(1997)는 보다 실용적인 검색어를 통해서 검색되어진 결과물에 대한 적합성 및 정확성을 고려하여 보다 개선된 검색 에이전트에 관한 연구를 하였다[2]. Leighton & Srivastava(1997)는 이러한 부분을 보완하여 연구발표 하였다[3]. 그러나 각 검색엔진에 대한 정의가 불분명하며, 키워드 및 디렉토리의 명확한 분류와 수집된 자료에 대한 정확한 통계적 검증에 단점을 가지고 있다. Gordon & Pathak(1999)은 8 개의 검색엔진을 통하여 이러한 선행연구의 단점을 보완하였다. Bradlow & Schmittlein(2000)는 20 가지 마케팅용어와 6 가지 검색엔진을 통해 이들간의 관계를 비교 분석 모형을 제시한 바 있다. 이상의 선행연구를 정리하면 [표 1-1]과 같다.

[표 1-1] 검색 Agent 평가에 관한 기존 연구정리

연구자	연도	검색용어 수	검색엔진 수
Leighton, Srivastava	1995	15	5
Chu, Rosenthal	1996	10	3
Schlichting, Nilsen	1996	5	4
Gordon, Pathak	1999	33	8
Bradlow, Schmittlein	2000	20	6

그러나 검색용어가 지나치게 전문적이어서 일반적인 검색엔진의 평가에는 다소 문제점

[1] A system architecture for intelligent browsing on the Web, hsiangchu lai, Tzyy-ching Yang 2000. 5

[2] 2nd Annual search engine shoot-out. *PC Computing* (1997)

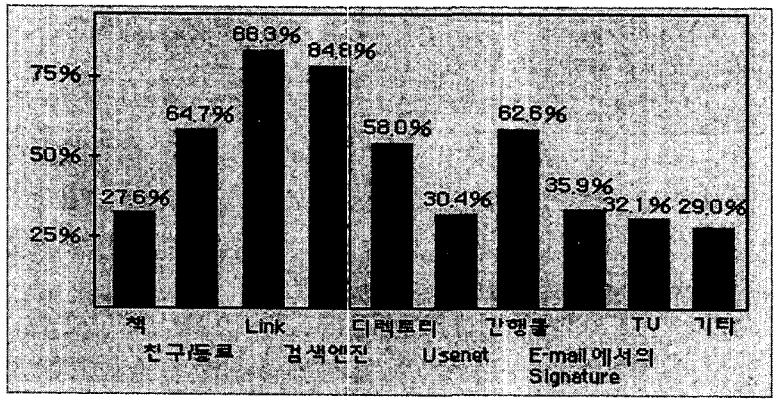
[3] Precision among World Wide Web (WWW) index services (search engines) : AltaVista, Excite,

이 있었다. 여기서는 이런 기존 연구들을 바탕으로 검색 에이전트의 평가를 위해 데이터베이스를 구축을 하였으며, 본 연구를 통해서 이전 연구의 문제점들을 개선하고자 한다.

2. 검색 엔진의 역할 및 구현

제 1 절 검색엔진의 역할

일반적으로 특정 사이트는 여러 가지 방법을 통해 알려지게 되며 접속하게 된다. [그림 2-1]의 조사결과를 통해서 보면, 이용자들이 특정 사이트를 발견하는 방법에는 배너광고나 타 사이트의 링크를 통한 접속이 가장 높았으며 키워드 검색을 제공하는 검색엔진과 특정 영역별로 구분해 놓은 디렉토리를 통해 찾아가게 되는 비율이 각각 84.8%와 58%로 나타났다. 즉, 검색엔진이나 디렉토리는 링크 다음으로 특정 사이트를 발견하는 도구로 이용되고 있음을 나타낸다.



출처 : searchenginewatch.com

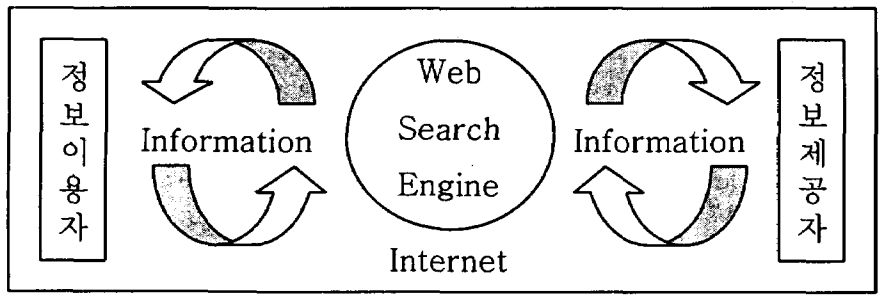
[그림 2-1] 인터넷에서 특정사이트를 발견하는 방법(중복허용)

링크와 검색엔진 및 디렉토리를 제외하면 친구나 동료로부터 소개 받아 웹사이트를 찾는 구전(Word of Mouth)의 비율이 64.7%였고, 신문, 잡지 등의 간행물을 통해 알게 되는 비율이 62.6%를 차지하였다. 한편 TV 광고나 Usenet, E-mail 의 마지막 부분에 일반적으로 넣는 Signature 가 차지하는 비중은 약간 낮은 것으로 조사되었다. 95 년도(제 3 차 조사)부터 시작된 이 조사는 이후에 큰 변화가 없이 링크와 검색엔진/디렉토리를 통한 방법이 웹사이트를 찾아가는데 가장 중요한 역할을 하는 것으로 조사되었다.

초기 인터넷 사용자들은 원하는 정보를 얻기 위해서 직접 인터넷을 돌아다니며 얻을 수 밖에 없었다. 그러나, 점차 인터넷 이용자 및 정보제공 사이트들이 증가하면서 이런 노

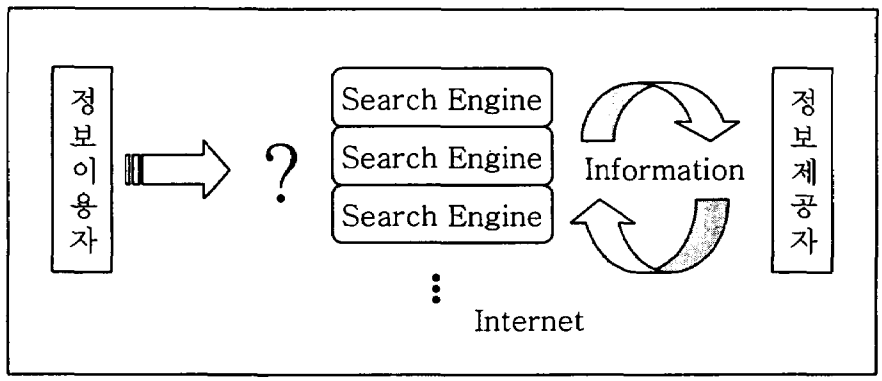
Hotbot, Infoseek, Lycos.(1997)

력을 대신해 줄 수 있는 검색엔진이 등장하게 되었다.



[그림 2-2] Internet 에서 검색엔진의 역할

[그림 2-2]는 정보의 바다라고 불리 우는 인터넷에서의 검색엔진의 역할이다. 검색엔진은 인터넷에서 정보의 수요, 공급자 사이의 연계사슬 역할을 한다. 그러나 점차 검색엔진의 수가 증가하면서 [그림 2-3]과 같이 이용자들은 어떤 검색엔진을 사용해야 하는가? 라는 문제점이 발생하게 된다.



[그림 2-3] 검색엔진 선택에 대한 문제발생

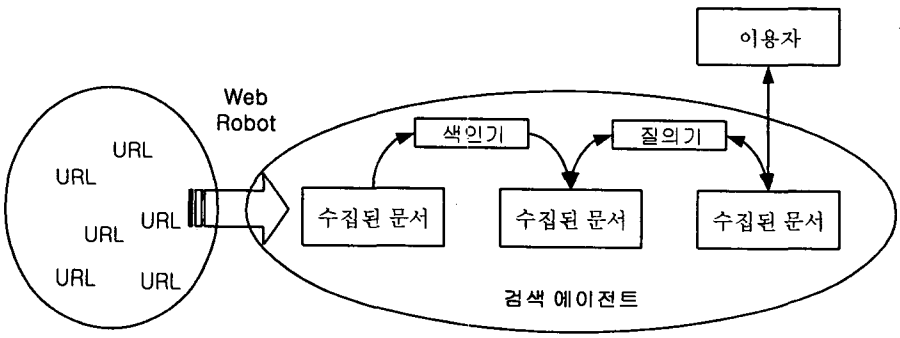
제 2 절 검색엔진의 구현 방식

검색엔진은 구현 유형에 따라 Web-Directory 방식과 Web Index 방식 그리고 메타형 검색방식으로 구분된다. 첫째, Web-Directory 방식은 검색사이트의 전문적인 정보검색사들이 직접 웹사이트들을 돌아다니며 정보를 수집해 오는 방식이며 둘째, Web-Index 방식은 사람에 의한 정보수집이 아닌 검색 에이전트인 로봇을 이용해서 정보를 수집하는 방식이다. 셋째, 메타형 검색방식은 자체검색기능을 보유하지 않고 여러 가지 검색엔진의 결과를 수집해서 이용자에게 보여주는 방식을 말한다. 이를 정리하면 [표 2-1]과 같다.

[표 2-1] 검색 엔진의 구현 유형에 따른 분류

종류	운영 메커니즘	장· 단점	비고
Web-Directory 방식	인터넷 검색사들이 직접 웹 사이트들을 돌아다니며 정보 수집→ 문서 주소를 주제별. 계층별로 디렉토리화	*키워드를 몰라도 몇 가지 분류만 알면 쉽게 정보에 접근 가능 *'98 년부터 사이트 수와 정보량이 급증하면서 방법적 한계에 직면	*야후의 첫 서비스 방식 *대분류→중분류→소분류를 따라 검색 *현재, Web-Index 방식과 병행해 사용됨
Web-Index 방식	검색 로봇이 주기적으로 인터넷에 있는 문서정보들을 검색→이를 자체 호스트 컴퓨터에 보내 Index DataBase 구성→사용자가 키워드를 치면 이에 해당하는 웹문서를 검색	*대량의 정보 처리 *데이터베이스에 저장된 것이 출력되므로 결과물의 대부분이 15 일에서 1 개월 전의 것임 *검색된 문서량이 너무 많고 그 중복도가 심함	*Altavista, Infoseek, Lycos(이상 미국), Naver, Simmani, Hanmir(이상 한국) 등 대부분의 검색서비스는 웹 인덱스 방식을 기본으로 구축되어 있음
메타형 검색방식	자체적인 검색 기능을 갖고 있지 않지만 여러 검색엔진과 연계되어 있어, 이들을 검색함으로써 결과를 한데 모아놓고 사용자로 하여금 선택하도록 함	*검색엔진에 따라 Data Base 가 다르기 때문에 검색결과가 다른 것을 활용할 수 있음	*한국의 와카노(Wakano) 는 메타 검색방식과 실시간 자동분류 시스템을 접목해 독일, 일본에서 서비스 시작

현재 대부분의 검색 서비스는 Web-Index 와 Web-Directory 방식을 병행한다. Web-Index 방식을 구현하는 검색엔진은 Web Robot 과 검색 에이전트의 두 부분으로 구성되어 있다.



[그림 2-4] 검색엔진의 구성 모듈

Web Robot 은 보통 웹 스파이더(Web Spider), 혹은 웹 크로울러(Web Crawler)라고도 얘기하는데 웹 상에서 자동으로 문서를 수집하는 역할을 한다. Web Robot 은 웹서버를 순회하면서 각 홈페이지에 있는 수많은 정보를 수집하는 프로그램을 말하며, 수집한 문서를 분석해 그 안의 URL 을 추출하여 다른 URL 을 연결시켜 주는 기능을 가지고 있어 사용자가 웹사이트를 돌아다니면서 웹 문서를 수집하지 않아도 자동으로 문서를 수집한다. 이렇게 모아온 문서를 바탕으로 색인기가 색인을 한다.

검색 에이전트를 구성하기 위해서는 원문 문서에 대한 색인 목록을 작성해야 하는데,

색인 목록은 원문을 분석해서 검색 키워드를 추출해 찾기 쉬운 형태로 내용을 기록해 놓은 것을 말한다. 이렇게 색인된 문서를 사용자의 질의어에 알맞게 선택해 보여주며 다양한 옵션을 통해 이용자의 복잡한 요구사항에 대응할 수 있게 된다.

3. 연구 수행 방법 및 설계

제 1 절 연구수행 개요

연구목적 달성을 위하여 우선 기존의 연구를 바탕으로 다음과 같은 단계로 분석하고자 한다. 우선 첫번째 단계로, 분석을 위한 데이터베이스의 구축을 위하여 검색엔진에 입력할 키워드를 선정 한다. 본 연구에서는 각 검색사이트에서 발표한 네티즌 인기 검색어 중에서 비교적 유행에 민감하지 않은 11 개의 키워드를 선택하였다. 두 번째 단계는 실험 대상인 검색 사이트의 선정이다. 국내 이용자 설문조사를 통하여 가장 선호하는 알타비스타(AltaVista), 한미르(Hanmir), 라이코스(Lycos), 네이버(Naver), 심마니(Simmani) 다섯 곳을 본 연구에선 선정하였다. 세번째 단계로는 실험 검색 사이트에 선택된 키워드를 검색해서 나온 결과 URL(Uniform Resource Locator) 데이터등으로 데이터베이스를 구축하였다. 각 실험 검색 사이트는 구현 방식에 따라 Web-Index 방식과 Web-Directory 방식으로 분류하여 데이터가 수집 되었다. 끝으로 데이터 분석을 통하여 각 실험 검색 사이트들의 유사성 파악 및 평가를 하였다.

제 2 절 키워드(Key-Word)의 선정

[표 3-1] 인기 검색어 100

한게임	삼국지 7	신화	포켓몬스터	사랑
애뮬	GOD	SBS	리니지	악보
세이클럽	박지윤	스포츠서울	디아블로 2	배경화면
아이러브스쿨	일간스포츠	일본게임	논문	성인만화
철권	삼국지	다모임	영화	스타크래프트
지도	쇼핑몰	보아	한메일	디아블로
삼성	스포츠조선	MBC	해적	가을동화
조선일보	게임자료실	KBS	해킹	동영상
검색엔진	네이버	경실연	아이콘	성인
모교사랑	성인영화	섹시	캐릭터	네오지오
서태지	만화	다음	미스코리아	고전게임
미소녀	성인방송	MP3	십자수	정품
에로	조성모	노란국물	가요	무료영화
메탈슬러그	포트리스	채팅	엽기	게임
펭클	WAREZ	자료실	구슬기	정품게임
유틸리티	이미지	부동산	HOT	안전보호구
홈페이지	아르바이트	와레즈	엽기토끼	레포트
독후감	핸드폰	다운로드	바람의나라	안티
리포트	다운	뮤직비디오	사진	시
카드	바탕화면	음악	유틸	일본

자료 : 네이버

연구에 필요한 URL 데이터베이스 구축을 위해 사용될 키워드를 선정 해야 했다. 국내 검색 사이트 중 네이버(http://www.naver.com)와 라이코스(http://lycos.co.kr)에서는 이용자들이 선호하는 검색용어를 정기적으로 공시하고 있으며, 본 연구에서는 네티즌들이 선호하는 키워드를 사용하기로 하였다. 인기검색어 중 다음과 같은 사항의 용어는 제외하기로 했다.

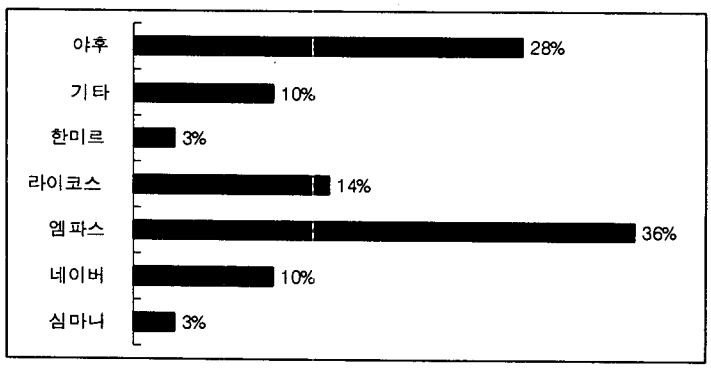
1. 지나치게 유행에 민감한 검색어
2. 특정 사이트나 연예인과 같은 특정인을 찾는 검색어
3. 검색결과 URL 수가 극히 작은 경우

이러한 선별 작업을 거쳐서 네이버의 인기검색어 100 개중에서 9 개를 선택하였고, 라이코스 인기검색어 50 개 중에서 ' 운세' , ' 유머' 두 가지를 합하여 총 11 개의 키워드(검색엔진, 논문, 만화, 부동산, 쇼핑몰, 영화, 운세, 유머, 지도, 채팅, MP3)를 사용하기로 하였다.

제 3 절 검색사이트의 선정

포탈 검색사이트 중에는 카테고리 방식과 키워드방식이 있다. 검색엔진의 선정은 키워드 선정과 마찬가지로 네티즌에게 인기 있는 포탈 검색사이트 중에서 엔진 유형이 주로 키워드 방식인 경우를 선정하였다. 카테고리방식은 에이전트로부터 수집된 정보를 검색엔진 관리자가 선별함으로써 검색 에이전트의 평가에 부적절한 것으로 판단되어진다. 이러한 이유로 카테고리방식인 야후와 엠파스는 실험 검색사이트 선정에서 제외 하였다.

[그림 3-1]은 인터넷 설문기관 ' Iloveinfo' 에서 조사한 설문 결과로 100 명의 전문 패널을 대상으로 가장 선호하는 검색엔진이 무엇인가를 조사한 것이다. 이와 같은 설문 자료를 바탕으로 키워드 방식의 에이전트를 사용하고 있는 알타비스타, 한미르, 라이코스, 네이버, 심마니를 실험 검색 사이트로 선정하였다.



출처:http://iloveinfo.co.kr

[그림 3-1] 검색사이트 순위 (설문 투표)

제 3 절 자료 수집 및 데이터베이스구축

검색엔진별 데이터의 구조는 출력순위, URL, 깨진 링크, 중복체크, 관련 디렉토리의 필드로 구성하였다.

[표 3-2] 알타비스타-'채팅'의 수집자료형식

검색순위	URL	깨진링크	중복체크	디렉토리분류
1	http://www.webboy.co.kr			12
2	http://nowplus.com			15
3	http://wons.com.ne.kr			2
4	http://155.230.19.124/chat3/mcClient.html	1		15
5	http://210.92.188.164	1		15
6	http://chat.miss4u.com			15
7	http://chat.now21.net/			15
8	http://chat.taejonnet.co.kr			15
9	http://fancyria.com			15
10	http://ice.hannam.ac.kr/~i2450027	1		15
11	http://muhanchat.net	1		15
...	...	...	...	...
38	http://westwood.fortunecity.com/moschino/209/			15
39	http://inhavision.inha.ac.kr./~s964568/wow.htm	1	184	15
...	...	...	...	...
184	http://inhavision.inha.ac.kr./~s964568/wow.htm	1	39	7

※ A:깨진링크(1:깨진링크), B:자체중복 발생시 중복되는 URL 의 검색순위 표시, C:검색엔진별 분류디렉토리

[표 3-2]는 알타비스타에서 ' 채팅' 키워드를 질의 해서 얻은 데이터의 일부이다. 검색엔진에서 매겨지는 출력 순위로 1 에서 10 번까지는 한페이지의 자료이며 흔히, 우리가 검색엔진에서 보는 첫 장을 말한다. 깨진링크는 LinkBot 을 이용하여 응답없는 URL 의 경우 1 로 체크하였다. 중복체크는 자체검색사이트의 결과내에서의 중복 URL 을 체크하였으며 표의 39 번 URL 과 184 번의 URL 이 중복되므로 중복되는 검색순위 번호를 교환입력하였다. 디렉토리 분류는 검색사이트별로 각각 다른 기준으로 분류되는 디렉토리를 20 가지로 통합한 후 그 번호를 기입하였다.

같은 키워드의 검색물의 URL 이 다른 디렉토리로 분류되는 것을 알아보기 위해 [표 3-3]과 같이 각 검색엔진의 디렉토리를 비슷한 속성끼리 분류하는 작업을 하였다. A 는 뉴스/미디어에 관련된 속성이며 B 는 비즈니스/경제에 관련된 속성을 부여했다. 이런 속성별로 키워드의 디렉토리 분포를 알아보았다.

[표 3-3] 각 검색엔진별 디렉토리 분류

	네이버	라이코스	한미르	심마니	알타비스타	속성
1	뉴스,미디어	뉴스, 미디어	뉴스,미디어	뉴스,언론	뉴스/미디어	A
2	비즈니스,경제	비즈니스, 경제	비즈니스,경제	비즈니스,경제	경제/금융	B
3			기업,회사	기업,회사		B
4	쇼핑	쇼핑		쇼핑		C
5	가정,여성	생활, 여성	생활,가정		여성/생활	C
6	사회,문화	문화, 사회	사회,종교			D
7			인문,사회과학	사회,생활	사회/문화	D
8	학문,과학	과학, 학문	과학,기술			D
9	교육,참고자료	교육,참고자료	교육,취업	교육,학습		E
10		취업정보		학문,참고자료	교육/취업	E
11			정보,공공기관	정치,행정		K
12	엔터테인먼트,예술	예술, 엔터테인먼트	문화,예술	엔터테인먼트	엔터테인먼트	G
13			연예,오락	예술		G
14	게임		취미,개인홈페이지	게임		G
15	컴퓨터,인터넷	컴퓨터, 인터넷	컴퓨터,인터넷	컴퓨터,인터넷	컴퓨터/인터넷	H
16	레크리에이션	레크리에이션				I
17	스포츠	스포츠	여행,레저스포츠	취미,스포츠	레저/스포츠	I
18	건강,의학	건강, 의학	건강,의학	건강,의학		I
19	지역정보	국가정보, 지역정보	지역정보	지역정보		D
99	기타	기타	기타	기타	기타	K

4. 분석

제 1 절 키워드와 디렉토리간의 관련성 분석

우선, 선정된 키워드로 수집된 자료의 디렉토리 속성별 분포를 알기 위하여 9 개로 분류한 디렉토리에 속하는 URL 수를 도출하였다. 그리고, 검색된 총 URL 수로 각 검색사이트에 키워드와 디렉토리간의 관련도를 도출하였다. 관련도의 도출은 (식 4-1)과 같다.

$$r = \frac{u}{U} \times 100 \qquad \text{(식 4-1)}$$

r: 관련도(%), u: 해당 디렉토리에 속하는 URL 수, U: 검색된 총 URL 수

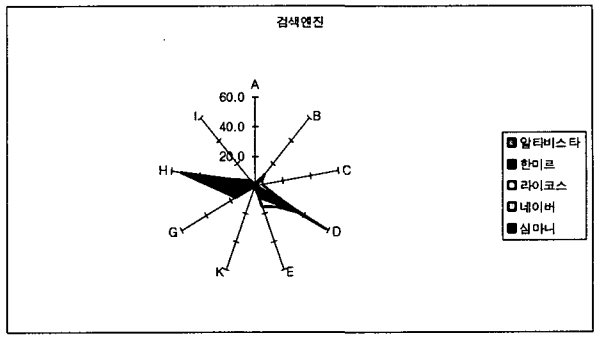
[표 4-1]은 ' 검색엔진' 을 키워드로 하였을 때 각 속성별 관련도(%)를 나타낸다.

[표 4-1] '검색엔진'의 디렉토리 분포

검색엔진	A	B	C	D	E	K	G	H	I
알타비스타	0.8	9.7	3.7	36.9	12.4	0.0	7.2	27.5	1.7
한미르	1.2	9.2	1.8	12.2	3.2	0.2	16.0	53.2	3.0
라이코스	0.8	8.0	5.2	29.3	15.0	0.0	4.8	32.1	4.8
네이버	0.2	11.0	1.0	58.4	1.4	0.0	4.0	22.8	1.2
심마니	0.2	7.6	0.8	36.2	8.4	0.6	4.4	40.0	1.8

주1 : A~I 는 표 의 각 검색엔진별 디렉토리 분류에 준함 / 주2 : 단위 : %

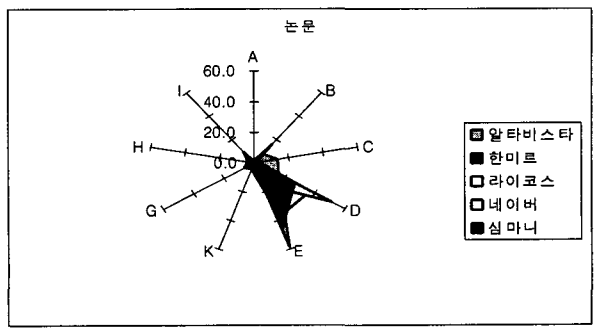
알타비스타의 경우 ' 검색엔진' 을 키워드로 사용한 결과 사회/과학(D)이 36.9%로 관련성이 가장 높게 나타났으며 컴퓨터/인터넷(H) 또한 27.5%로 두 번째로 높은 관련성을 보였다.



주 : A~I 는 표 의 각 검색엔진별 디렉토리 분류에 준함

[그림 4-1] '검색엔진'의 디렉토리 분포

[그림 4-1]은 ' 검색엔진' 을 키워드로 입력한 결과와 URL 의 디렉토리 속성별 관계도 분포를 나타내고 있다. A 에서 I 까지의 축은 키워드별 검색엔진의 디렉토리 속성별 분포를 나타낸다. 키워드 ' 검색엔진' 의 경우는 ' 컴퓨터/인터넷' 과 ' 사회/과학' 으로 분류되는 경우가 많이 나타났다. [그림 4-2]와 같이 ' 논문' 의 경우는 ' 교육/취업' 과 ' 사회/과학' 으로 분류 되는 경우가 높았으며, 알타비스타의 경우 생활/쇼핑 속성이 약간 나타났다.



주 : A~I 는 표 의 각 검색엔진별 디렉토리 분류에 준함

[그림 4-2] '논문'의 디렉토리 분포

이와 같은 방법으로 분석한 결과로 ' 만화' 는 사회/과학, 엔터테인먼트/예술, 컴퓨터/인터넷으로 세 종류의 관련도가 높은 것으로 나타났다. ' 부동산' 은 비즈니스/경제와 사회/과학의 관련도가 높은 것으로 나타났으며, ' 쇼핑몰' 은 한미르가 비즈니스/경제 심마니가 생활/쇼핑에 관련도가 높은 분포를 나타냈다. ' 영화' 는 엔터테인먼트/예술부분과 사회/과학부분에 관련도가 높게 나타났다. ' 운세' 는 사회/과학, 엔터테인먼트/예술, 비즈니스/경제 레포츠/건강의 순으로 다양한 관련도 분포를 보였다. ' 유머' 의 디렉토리는 엔터테인먼트/예술과 사회/과학에서 높은 관련도 분포를 나타냈으며 라이코스는 레포츠/건강에 약간의 URL 이 나타났다. ' 지도' 는 사회/과학과 교육/취업부분에 높은 관련도분포를 보였으며 라

이코스는 컴퓨터/인터넷에 분포를 나타냈다. ' 채팅' 의 분포는 ' 지도' 와 마찬가지로 다양한 분포를 나타냈다. 엔터테인먼트/예술과 컴퓨터/인터넷 그리고 사회/과학부분에 높은 관련도 분포가 나타났다. 끝으로 ' MP3' 는 엔터테인먼트/예술과 사회/과학부분에 높은 관련도 분포를 보였다.

Web-Directory 방식에서는 전체적으로 컴퓨터, 인터넷에 분류되어지는 URL 이 가장 많은 것으로 나타났다. 다음으로는 사회, 문화, 종교, 인문, 과학, 기술, 생활이 높게 나타났다.

[표 4-2] 키워드와 디렉토리간의 관련성

키워드	디렉토리 분류 코드
검색엔진	컴퓨터, 인터넷, 사회, 문화, 인문, 과학, 기술, 생활
만화	상동, 엔터테인먼트 추가
채팅	상동
영화	사회, 문화, 인문, 과학, 기술, 생활, 엔터테인먼트
운세	상동
유머	상동
MP3	상동
쇼핑몰	쇼핑, 생활 여성
지도	교육, 학습, 참고자료, 취업, 사회, 문화, 인문, 과학, 기술, 생활
논문	상동
부동산	비즈니스, 경제, 금융, 기업, 회사

각 검색 에이전트의 키워드 검색결과에서 분류 되어지는 디렉토리를 이용하여 키워드와 디렉토리간의 관련성을 살펴보면, 먼저 키워드 ' 검색엔진' 과 ' 만화' , ' 채팅' 은 컴퓨터, 인터넷에 주로 분류되며, 쇼핑몰은 비즈니스, 경제에 분류되는 성향도 있음을 알 수 있다. 그밖에 실험 대상인 키워드들은 사회, 문화, 종교, 인문, 과학, 기술, 생활에 분류 됨을 알 수 있다. 키워드간의 디렉토리의 유사 집단으로 분류하면 [표 4-2]와 같다.

제 2 절 검색 에이전트의 유사성 분석

수집된 URL 을 이용해서 같은 검색 키워드에 대한 검색 에이전트별 특성을 알아보았다. 먼저, 몇 가지 코딩작업이 필요 하며, 그 작업은 다음과 같다. 우선 자체 중복검사로서 단일 검색엔진의 결과 내에서 동일한 URL 이 반복적으로 나타나는 경우, 이런 중복 URL 을 제거하는 작업을 시행하였다.

두 번째 작업으로는 검색된 URL 의 깨진링크(Broken Link) 여부검사를 실시하였다. 검색 에이전트중에는 특별히 이런 깨진링크를 방지하기 위해서 링크검사 에이전트가 각 검색엔진별로 존재한다. 이와 같은 깨진링크 검사를 통해 각 검색엔진별 링크 검사 에이전트를 평가를 할 수 있다. 앞에서 선택한 키워드를 검색엔진에 질의한 후 출력된 URL 을 수집하여 키워드별, 검색엔진별로 정리하였으며, 그 결과는 [표 4-3]과 같다. 검색되어진 결과물의 양, 즉 수집 에이전트의 성능적 측면으로는 ' 심마니 > 알타비스타 > 네이버 > 한미르 > 라이코스' 순으로 나타났음을 볼 수 있다.

[표 4-3] 검색엔진별 검색결과 수

	알타비스타	한미르	라이코스	네이버	심마니	합계
MP3	2942	2842	3527	1863	5053	8232
영화	6097	3384	2432	3571	5956	9387
채팅	1827	913	588	928	1313	2429
만화	2193	1409	984	1227	3224	3620
지도	1674	910	427	930	1216	2267
부동산	3064	1940	1971	2139	3415	6050
쇼핑몰	4838	3390	3523	4513	4648	11426
논문	2135	1034	640	1051	1547	2725
가요	1240	1055	517	756	1402	2328
검색엔진	2826	936	1069	1497	1959	3502
유머	2156	925	1214	1388	2923	3527
운세	525	213	256	325	570	794
합계	31517	18951	17148	20188	**33226**	56287

[표 4-4] 키워드-검색엔진별 중복된 URL 의 수

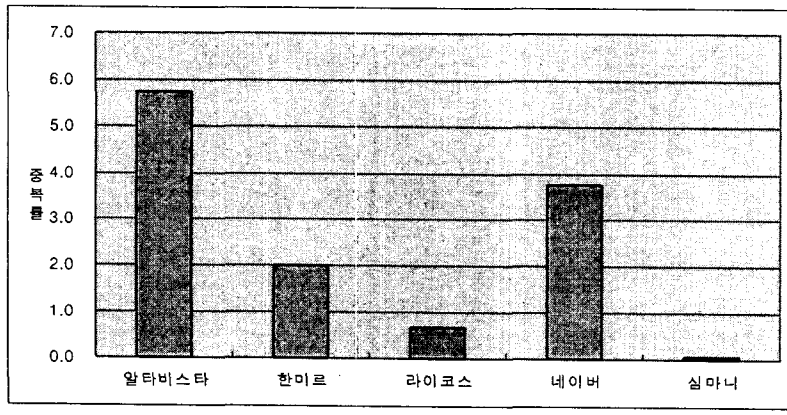
	알타비스타		한미르		라이코스		네이버		심마니	
	검색수	중복수	검색수	중복수	검색수	중복수	검색수	중복수	검색수	중복수
검색엔진	500	17	500	0	500	1	500	0	500	0
논문	500	27	500	0	500	23	500	0	500	0
만화	500	24	500	0	500	0	500	0	500	0
부동산	500	104	500	47	500	2	500	0	500	0
쇼핑몰	500	63	500	63	500	8	500	2	500	2
영화	500	0	500	0	500	0	500	0	500	0
운세	500	19	500	0	246	0	323	1	500	0
유머	500	9	500	0	500	1	500	1	500	1
지도	500	29	500	0	433	0	500	0	500	0
채팅	500	14	500	0	500	2	500	0	500	0
MP3	500	10	500	0	500	0	500	203	500	0

그러나, 검색 에이전트의 평가는 단순히 이런 결과로 평가할 수 없다. 이용자들의 85%가 검색결과의 첫페이지(즉, 검색순위 1~10 번)만 본 후에 검색작업을 마친다는 통계가 있으며, 따라서 본 연구에서는 수집된 데이터 중 검색순위 500 개까지만 이용하기로 하였다. 여기서, 라이코스와 네이버 등은 ' 운세' 와 ' 지도' 에서 검색결과 URL 이 500 개 미만으로 출력되는 경우도 있었다. [표 4-4]는 키워드-검색엔진별 중복된 URL 의 수를 나타낸다.

중복 URL 의 검사로 각 검색엔진별 키워드별 평균적인 중복률을 얻은 결과는 [표 4-5]와 같다. 네이버 ' MP3' 의 경우는 500 개 중에서 203 개가 중복되었으며, 그 외의 키워드에서는 매우 낮은 중복을 가지고 있었다. 심마니는 전체적으로 자체 중복이 가장 낮게 나타났다.

[표 4-5] 키워드-검색엔진별 중복률

	알타비스타	한미르	라이코스	네이버	심마니	평균중복률
검색엔진	3.4	0	0.2	0	0	0.7
논문	5.4	0	4.6	0	0	2.0
만화	4.8	0	0	0	0	1.0
부동산	20.8	9.4	0.4	0	0	6.1
쇼핑몰	12.6	12.6	1.6	0.4	0.4	5.5
영화	0	0	0	0	0	0.0
운세	3.8	0	0	0.3	0	0.8
유머	1.8	0	0.2	0.2	0.2	0.5
지도	5.8	0	0	0	0	1.2
채팅	2.8	0	0.4	0	0	0.6
MP3	2	0	0	40.6	0	8.5
평균 중복률	**5.7**	2.0	0.7	3.8	0.1	



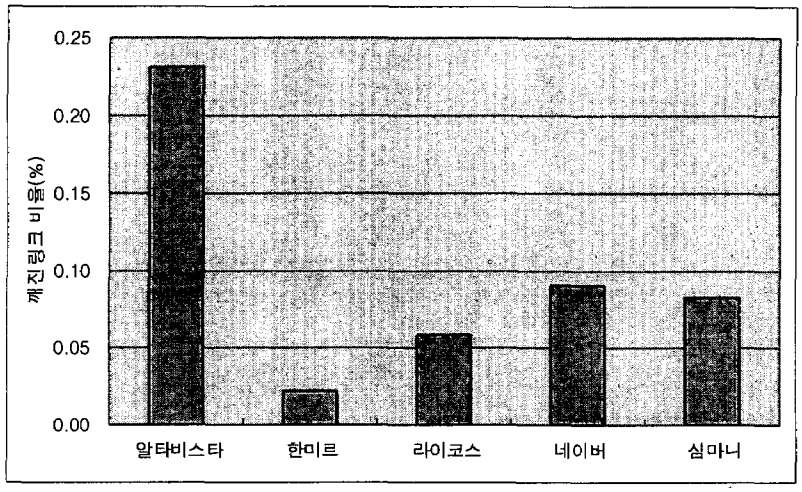
[그림 4-12] 검색엔진별 중복률 비교

[그림 4-12]에서 알 수 있듯이 알타비스타가 가장 높은 자체 중복률을 가진다는 것을 알 수 있다. 자체중복률은 같은 URL 을 중복해서 보여주기 때문에 검색작업에 비효율적 요소로 작용하게 된다.

[표 4-6] 검색엔진/키워드별 깨진 링크 비율

	알타비스타	한미르	라이코스	네이버	심마니
검색엔진	21.5%	3.4%	8.4%	9.8%	8.6%
논문	22.6%	2.6%	10.3%	6.6%	10.0%
만화	17.4%	5.6%	3.0%	8.4%	3.0%
부동산	32.6%	0.7%	10.0%	11.6%	8.2%
쇼핑몰	28.1%	1.2%	9.1%	9.4%	10.6%
영화	25.2%	1.0%	2.8%	9.0%	8.4%
운세	16.2%	3.0%	4.5%	7.8%	7.6%
유머	19.1%	1.2%	4.2%	7.4%	7.8%
지도	25.3%	1.4%	2.1%	8.8%	9.0%
채팅	24.9%	1.8%	5.2%	11.8%	8.2%
MP3	21.0%	2.4%	4.2%	8.8%	9.4%

검색엔진별 평균 중복률로 보면 ' 알타비스타 > 네이버 > 한미르 > 라이코스 > 심마니' 순으로 나타났다. 깨진링크를 찾기 위해서 ' Watchfire, LinkBot' [4]을 이용하여 검사를 실시하였다. [표 4-6]은 URL 중에서 깨진링크의 비율을 나타낸 것이다. 깨진링크의 비율에서도 역시 알타비스타가 가장 많은 깨진링크를 보였으며 한미르는 가장 낮은 깨진링크를 보유한 것으로 나타났다. 깨진링크의 빈도로 본다면 ' 알타비스타 > 네이버 > 심마니 > 라이코스 > 한미르' 순으로 성능이 높은 것으로 나타났다.



[그림 4-13] 검색엔진별 깨진링크(Broken-Link)비율

앞의 두 가지 작업을 통해 여과되어 나온 자료를 이용해 각 검색 에이전트간의 관계를 알아보았다. 5 개의 검색엔진을 이용해서 얻을 수 있는 조합의 수는 총 32 가지이며, 각각의 검색 사이트들간의 조합을 다음과 같은 벡터 형식으로 나타내었다.

$$(S_1, S_2, S_3, S_4, S_5)$$

$$S_n = \begin{cases} 1 & \text{해당 사이트 사용} \\ 0 & \text{해당 사이트 사용 안함} \end{cases}$$

S_1 = 알타비스타, S_2 = 한미르, S_3 = 라이코스, S_4 = 네이버, S_5 = 심마니

S_n 은 0 과 1 로 사이트 사용 유무를 나타 냈다. 예를 들어서 (1, 0, 0, 0, 1)은 알타비스타와 심마니를 이용한 검색결과의 중복 URL 수를 나타낸다. 검색엔진간 중복 URL 수를 알아본 결과는 [표 4-7]과 같다.

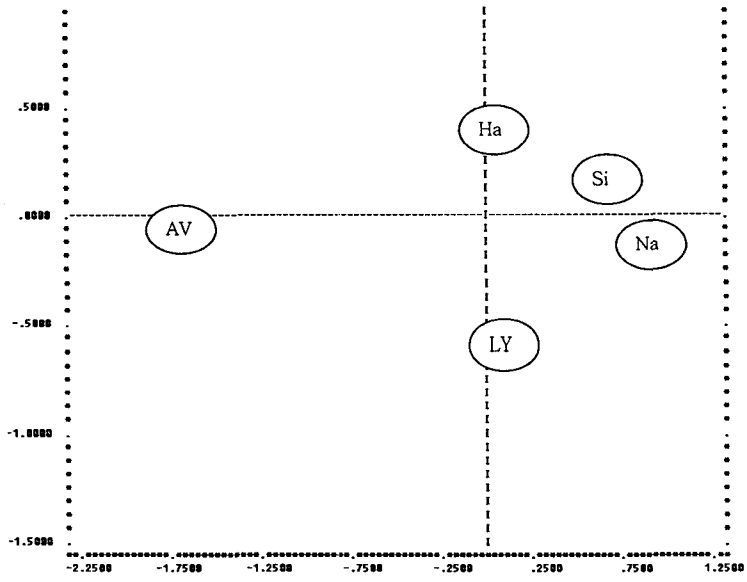
[4] Broken link 검색 소프트웨어

[표 4-7] 검색엔진간의 중복관계

		검색엔	논문	만화	부동산	쇼핑몰	영화	운세	유머	지도	채팅	MP3
1	1,1,0,0,0	34	37	39	25	26	35	24	15	21	18	10
2	1,0,1,0,0	21	42	18	30	36	10	27	18	24	17	5
3	1,0,0,1,0	36	37	81	51	42	54	50	32	32	39	17
4	1,0,0,0,1	14	17	0	26	28	24	69	37	16	32	31
5	0,1,1,0,0	49	38	52	60	46	32	41	43	47	86	56
6	0,1,0,1,0	52	43	66	103	88	58	56	31	41	60	17
7	0,1,0,0,1	3	2	0	4	0	2	2	3	5	0	1
8	0,0,1,1,0	41	29	39	75	33	26	41	17	44	63	16
9	0,0,1,0,1	3	4	0	3	21	6	9	5	9	3	5
10	0,0,0,1,1	8	33	0	17	3	35	23	16	18	7	13

※ 알타비스타, 한미르, 라이코스, 네이버, 심마니(자체중복률 제외)

검색 엔진들간의 중복수가 많을수록 비 보완관계에 있다고 보았으며, 반대로 중복이 적을수록 보완관계의 검색엔진이라고 보았다. 이와 같은 자료를 바탕으로 검색 사이트간의 관계를 파악하기 위해 다차원척도법(Multi-Dimensional Scaling)을 이용하여 분석하였다. KYST[5]를 이용하여 결과로 [그림 4-14]를 구하였다.



주 : Av=알타비스타, Ha=한미르, LY=라이코스, Na=네이버, Si=심마니, stress value= 0.008

[그림 4-14] 다차원척도법을 이용한 검색엔진들간의 관계

[5] Kruskal, Young, Shephard, Torgerson; 다차원척도법 프로그램

결과의 스트레스값(stress value)은 0.008 이다. 분석에 사용된 MDS 는 입력되어진 데이터가 중복률이다. 검색 엔진들간의 중복이 낮으면 가까이 위치하며 반대로 중복률이 높으면 상대적으로 거리가 멀게 된다. [그림 4-14]를 보면 알타비스타와 나머지 검색엔진들 간의 거리가 멀게 나타났다. 이는 알타비스타가 나머지 검색엔진들 간의 중복이 가장 높았다는 것을 말한다. 그리고 네이버와 심마니가 중복이 가장 낮다는 것을 알 수 있으며, 만약 네이버로 검색을 해서 만족할만한 검색결과를 얻지 못한 사용자는 심마니나 라이코스등을 이용하는 것이 중복되는 URL 의 수가 가장 적다는 측면에서 효율적인 검색이 될 수 있음을 알 수 있다.

5. 결 론

본 연구를 통해서 같은 키워드에 의한 검색결과도 각 검색 사이트의 에이전트에 따라 많은 차이가 있다는 것을 알 수 있었다. 먼저, Web-Directory 방식에서는 각 키워드가 주로 속하는 디렉토리 속성과 관계를 알 수 있었다. 그리고, 크게 분류되는 수를 5 가지로 축약할 수 있음을 볼 수 있었다. 전체적으로는 컴퓨터/인터넷에 분류되어지는 URL 이 가장 많은 것으로 나타났으며, 다음으로는 사회, 문화, 종교, 인문, 과학, 기술, 생활순으로 높게 나타났다.

둘째, Web-Index 방식에서는 각 검색 에이전트의 키워드 검색을 통한 중복되어지는 URL 의 빈도를 측정하여 각 검색 사이트간의 에이전트의 보완관계와 비 보완관계를 얻을 수 있었다. 이 같은 관계를 이용하여 검색엔진 사용자는 중복을 피할 수 있는 검색방법을 모색할 수 있다.

그러나, 이러한 결과는 동적인 인터넷 환경에서 실시간으로 변화한다.그러므로, 데이터의 수집과 분석 역시 실시간으로 행해져야 하며, 처리속도의 신속성이 요구된다. 따라서, OLAP 와 같은 실시간 분석이 가능한 데이터 수집 에이전트 및 분석 프로그램이 요구 된다. 또한, 본 연구는 다양한 키워드 중에서 11 개를 실험군으로 선택하여 연구 하였다. 추후, 추가 키워드에 관한 분석이 필요하다.

참고 문헌

국내문헌

[1]신봉기, 김영환 (1997), "웹 에이전트", 정보과학회지, 제15권 제3호, 61-67.

[2]이근배 (1998), "에이전트 기반 정보검색", 정보과학회지, 제16권 제8호, 32-38.

[3]이은석 (1997), "멀티에이전트 기술의 실세계 시스템으로의 응용", 정보과학회지, 제15권 제3호, 17-28.

[4]최중민 (1997), "에이전트의 개요와 연구방향", 정보과학회지, 제15권 제3호, 1-16.

국외문헌

[5] Bradlow, Eric T. and David C. Schmittlein (2000), "The Little Engines That Could: Modeling the Performance of World Wide Web Search Engines", *Marketing Science*, 19(1), 43-62.

[6] Feldman, S.(1997), "Just Answers, Please. Choosing a Web Search Service", ***http://www.infotoday.com/searcher/may/story3.htm.***

[7] Retrieval Effectiveness of Search Engines", *Information Processing and Management*, 141, 1999.

[8] Lake, M.(1997), "2nd Annual Search Engine Shoot-out", *http://www4. zdnet.com/pccomp /features/exc10997/sear/sear.html.*

[9] Leighton, & Srivastava (1997), "Precision among World Wide Web Search Services (search engines): AltaVista, Excite, HotBot, Infoseek, Lycos", *http://www.winona.msus.edu/is-f/library-f/webind2/webind2.htm.*

[10] Westera, G.(1996), "Robot-Driven Search Engine Evaluation Overview", *http://www.curtin.edu.au/curtin/library/staffpages/gwpersonal/senginestudy/index.htm.*