경영정보학연구
제13권 제1호
2003년 3월

메타 정보검색 에이전트 시스템을 위한 다중플랫폼의 데이터베이스 재구조화에 관한 연구

신 창 훈*, 류 병 무**

A Study on the Restructure of Multi-Platform Databases for MIRAS

Changhoon Shin, Pyungmu Ryu

It is important to retrieve information that a user requires on the web. The web is an open system. The amount of information is increasing rapidly. While each of information was compiled into the database piece at a single platform in the past, it is now compiled into complicated structure at a multi-platform. Restructuring the multi-platform database is needed to efficiently retrieve information. MIRAS (Meta Information Retrieval Agent System) has a multi-platform database on web. This study applies the classification of the existent site's categories to restructure the database systematically. The empirical analysis shows that the suggested method is effective for information retrieval and multi-platform database restructuring. This study helps users to save on time-cost of searching information.

* 한국해양대학교 물류시스템공학과 부교수
** (주) SLS 기획팀

Ⅰ. 서 론

웹(Web)에서 사용자가 필요한 정보를 검색하는 작업은 중요한 부분을 차지하고 있다. 이러한 웹의 체제는 개방된 형태이며, 정보의 양은 지속적으로 증가하고 있다. 과거에는 각 정보의 색인(Index)을 단일플랫폼(Single Platform)의 데이터베이스 형태로 구축하였으나, 오늘날에는 폭증하는 정보의 양에 따라 다중플랫폼(Multi-Platform)의 데이터베이스로 복잡하게 구축되고 있다. 이러한 환경에서 데이터베이스의 재구조화는 사용자가 요구하는 정보에 효율적으로 접근시키는데 중요한 부분을 차지한다.

본 연구는 기존의 검색 시스템들을 통합하는 메타 검색 에이전트 시스템[1]의 설계를 전제로 한다. 이 시스템은 기존의 각 검색엔진에서 제시하는 정보를 통합, 가공하여 사용자에게 보다 나은 서비스를 제공하는 것을 목적으로 하고 있다. 이러한, MIRAS의 대표적 요구사항을 정리하면 다음과 같다. 첫째, 사용자는 보다 편리하고 쉽게 원하는 정보를 검색할 수 있는 환경을 요구하며, 따라서 각종 정보검색서비스의 상호연계 및 통합 관리 기술과 함께 보다 새로운 정보검색서비스를 구축할 수 있는 기술이 요구된다. 둘째, 웹에서의 문서 및 사이트들은 많은 분산 데이터베이스로 존재한다. 이러한 분산된 웹 데이터베이스를 제공하는 각 검색 사이트들을 보다 효율적으로 통합 관리하는 방법이 요구된다. 셋째, 이러한 시점에서 보다 시스템 자원 및 비용을 효율화하는 MIRAS의 필요성이 대두된다. 현재, 이와 관련한 정보검색 시스템의 성능을 향상시키기 위한 연구는 계속되고 있다. 본 연구는 명확하지 않은 분류 환경에서의 일반메타정보검색, 쇼핑몰의 제품분류검색 등에 응용이 가능하며, 각종 데이터베이스의 효과적인 관리의 방안을 제시한다.

본 연구는 다중플랫폼의 데이터베이스를 체계적으로 재구조화하기 위해 각 검색 사이트들의 카테고리 분류를 응용한다. 먼저, 각 검색엔진에서 제시하는 카테고리를 수집하여 51개의 카테고리 단어(Category keyword)를 선정한다. 이를 근거로 단일 집단에서 51개의 카테고리명을 가지는 집단으로 분류할 수 있는 가능성을 전제로 연구가 수행되었다. 실험검색 대상인 키워드[2]는 앞서 51개의 카테고리 단어와 인기검색어 60개를 선정한다. 기존의 6개 검색사이트[3]에서 실험검색키워드에 따른 20개의 검색결과에 대한 사이트주소(Site URL), 제목(Title), 웹 디렉토리(Web Directory), 디렉토리 주소(Directory URL), 엔진명(Engine name), 설명(Description)을 수집한다. 수집된 데이터에서 실험검색키워드와 51개의 카테고리명들의 관련도(Degree of Relation)를 분석하고, 카테고리명 간의 관련도를 분석한다. 또한, 카테고리명 간의 상관관계행렬을 이용한 계보적 군집방법[4]을 응용하여 집단화 한다. 이 집단화에서 설명률[5]을 참조하여 다중플랫폼의 데이터베이스의 재구조의 대안을 설정한다. 이렇게 선정된 대안들을 기반으로 재구조화된 데이터베이스의 검색성능을 평가한다. 즉, MIRAS의 검색 재현률(Recall ability), 중복(Overlap), 소모시간(Time-cost)을 측정하고 관계를 규명하여 결과적으로 효과적인 다중플랫폼 데이터베이스의 재구조화 방안을 제시한다. 이러한 MIRAS의 구조는 은닉시스템[6]으로 존재한다. 또한, 실시간 데이터의 갱신과 학습을 통하여 최적의 분류 집단 수는 수정된다.

본문은 다음과 같이 구성된다. 먼저, 이론적 배경은 에이전트의 정의와 에이전트의 필요성

1) 이하, MIRAS(Meta Information Retrieval Agent System).

2) 이하 실험검색키워드.
3) 네이버, 라이코스, 심마니, 야후, 엠파스, 한미르.
4) Hierarchical Cluster Method.
5) Proportion Explained.
6) Cache System.

및 연구현황을 기술한다. 또한, 다중플랫폼의 메타검색에 대하여 기술하고, 그에 따른 데이터베이스의 선택문제를 알아본다. 여기서 문서의 분류 및 카테고리와 MIRAS의 정의, 데이터베이스의 구조를 살펴본다. 둘째, MIRAS의 실험설계 및 데이터베이스 구조 분석에 대하여 기술한다. 여기서 성능 평가와 사전(事前)카테고리구조분석, 실험검색키워드를 선별한다. 셋째, 데이터 수집 및 분석은 카테고리 간의 관계와 키워드와 카테고리간의 관계를 분석한다. 결과로 몇 개의 카테고리로 재구조화 되는 대안들을 만들어보고, 이때의 키워드와 재구조화된 카테고리들간의 관련도를 산출한다. 넷째, MIRAS의 검색 및 성능 평가를 한다. 결론에서 연구의 전반적인 시사점, 추후연구과제와 한계점을 서술한다.

Ⅱ 이론적 배경

2.1 에이전트

2.1.1 에이전트의 정의

사용자를 대신하여 원하는 작업을 자동적으로 해결해주는 소프트웨어들을 에이전트(Agent)라 하여, 많은 정보 기술의 영역에서 개발되어 왔다. 지금까지 에이전트란 정확한 정의를 내리지 못하고 있는 실정이며 대략적으로 사용자를 대신하여 유용한 작업의 일부를 실행하는 도구(Tool)로 의미를 내포하고 있다. 이러한 에이전트의 개념을 웹 환경에서 고객-서버의 대화 도구로 이용하고자 하는 연구가 1990년대 중반부터 활발히 진행되었다.

2.1.2 고객-서버 모형에서의 에이전트의 필요성

웹에서 분산처리는 매우 단순한 프로토콜 HTTP[7]와 RPC[8] 동작원리에 따라 이루어진다. 이 방식은 전통적인 고객-서버 모형에 따른다[신봉기, 김영환, 1997]. 이 모형은 고객과 서버간의 대화가 많이 필요한 경우 상대적으로 느린 통신처리시간이 문제점으로 나타난다. 따라서, 중간자인 에이전트의 개입으로 쌍방의 정보교류에 있어서 시간비용요인을 단축시키려는 시도가 진행되게 된다.

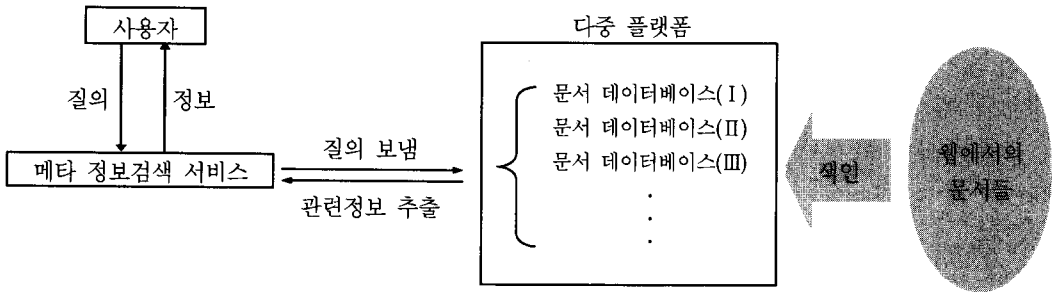
인터넷에서 정보제공서비스는 웹에서 분산된 지식과 정보들을 사용자의 요청에 의한 지정된 정보만을 전달하는 수동적인 성격을 가지고 있다. 이러한 환경에서 제각기 다른 목적을 갖고 스스로 동작하는 에이전트들을 구성함으로써, 능동적으로 성능을 향상시킬 수가 있다. 예를 들어, 사용자의 성향을 파악하는 에이전트, 검색키워드간의 관련도, 또는 그에 따른 관련문서의 상호정보를 파악하는 에이전트 등이 하나 이상으로 결합하여 정보검색에 개입함으로써, 그 성능향상을 모색할 수 있다. 특히, 에이전트에 기반 한 정보검색 시스템[9]은 각 에이전트의 특성과 해당 서비스의 종류에 따라 추천에이전트[10] [Gerald, Valerie, 1999], 뉴콘텐츠에이전트[11], 검색에이전트[12], 맞춤에이전트[13], 개인화에이전트[14] 등으로 구성하여 능력을 향상 시킨다[lai, Yang, 1998].

2.2 정보검색

2.2.1 단일플랫폼의 정보검색

전통적인 정보검색은 일반적으로 각 문서에 대한 색인을 단일플랫폼의 데이터베이스 형태로

7) Hyper Text Transfer Protocol.
8) Remote Procedure Call.
9) IRAS(Information Retrieval Agent System).
10) Recommendation Agent.
11) New-contents Agent.
12) Search Agent.
13) Customized Agent.
14) Personal-status Agent.



<그림 1> 메타검색서비스

구축하고, 이에 따른 키워드와 문서의 관련도는 $TF \times IDF$ 알고리즘에 기반한 벡터공간모델을 사용한다[Salton, McGill, 1983]. 여기서, 관련도 W_{ij}는 다음 식 (1)로 산출한다. TF_{ij}는 단어 T_i가 문서 D_j에 나타나는 빈도수이며, IDF_i는 단어가 등장하는 문서의 상대도수를 나타내는 DF_i의 역수에 로그변환을 한 값이다.

$$W_{ij} \equiv TF_{ij} \times IDF_i,$$
$$IDF_i \equiv \log(N/DF_i), \quad (1)$$
$$N \equiv \text{총 문서수}$$

즉, IDF_i는 전체문서 중, 어휘 i가 존재하는 문서의 비율을 의미한다. 이러한 방식은 라이코스[15]와 같이 사용자가 많은 경우 빈번한 검색으로 인한 심한 병목현상과 접근지연현상을 일으킬 수 있는 단점을 가지고 있다[Mauldin, Leavitt, 1994]. 사실상 인터넷에서 정보는 생성, 소멸, 변경 등이 빈번히 이루어지며, 이러한 정보를 단일 플랫폼에 일관적이며 효과적으로 관리하는 것은 불가능하다.

2.2.2 다중플랫폼의 메타 검색

단일플랫폼의 문제를 해결하기 위하여, 정보검색을 각자의 정보 문서에 대한 색인을 가지고 있는 문서 데이터베이스들에 분산시키는 방법이 출현하였다. <그림 1>은 메타검색 서비스[16]라 불리는 분산웹검색 에이전트를 말한다. 그림에서 다중플랫폼의 문서 데이터베이스들은 문서의 모임과 임의의 질의에 대하여 그 문서 모임으로부터 관련된 문서를 반환 해주는 색인시스템으로 구성된 정보자원을 뜻한다. 일반적으로 주제별검색엔진과, 문헌정보검색서비스 등은 모두 문서 데이터베이스이다. 이러한 분산된 다중플랫폼의 데이터베이스에서 정보검색을 수행하는 형태를 메타검색 서비스라 한다.

2.2.3 데이터베이스의 선택문제

문서 데이터베이스는 대부분 유사한 정도로 군집화 된다. 따라서, 사용자는 원하는 정보를 하나 또는 몇 개의 문서 데이터베이스에서 획득할 수 있다. 즉, 사용자의 질의에 관련 없는 문서 데이터베이스를 열람하는 것을 방지한다. 그러므로, 시스템 자원소모와 불필요한 통신 비용을 줄이며 효과적인 검색 작업을 수행할 수 있게 된다. 이와 같이 메타 정보검색 서비스에서는 다중플랫폼의 형태인 여러 개의 문서 데이터베이스를 가지며, 그것들을 선택하는 문제가 관심으로 대두 되었고 이러한 관련 연구가 진행되기 시작했다.

Bicchieri(1998)는 데이터베이스 선택문제[17]를

15) http://www.lycos.com.

16) Meta-Retrieval Service.

17) Database Selection Problem.

<표 1> 데이터베이스 선택문제의 선행연구

분 류	System	내	용
통계적 방법	SMART[a]	방 법	각 문서 데이터베이스의 중심점 용어벡터(Centroid Term-Vector)를 그 문서 데이터 베이스의 색인으로 사용하여 사용자의 질의에 유사도를 측정하는 방법을 사용.
		장 점	주제에 따라 각 문서데이터 베이스가 잘 분류되었을 때 유용함.
		단 점	각 용어의 분포 상황에 따라 실제 정보검색 환경에서 잘못된 결과를 보여주는 경우도 있음.
	Gloss[b]	방 법	각 용어에 대한 문서 빈도나 용어 가중치 합과 같은 통계적 정보를 기반으로 하여 사용자 질의에 대한 각 문서 데이터베이스의 관련도를 계산하는 방법을 사용.
		장 점	특정 환경에서 매우 효과적임.
		단 점	매우 제한적인 가정들을 기반으로 하여 만족하지 않는 많은 검색환경에서 응용의 제한이 있음.
증강 학습방법	Savvy[c]	방 법	사용자 질의에 대한 정보검색 결과들을 기록하고, 훈련 데이터로 사용하여 각 문서 데이터 베이스와 질의 용어의 관련도를 증강학습법으로 구하고 이를 바탕으로 임의의 질의에 대한 각 문서 데이터 베이스의 관련도를 계산하는 방법을 사용.
		장 점	사용자에 대한 적응성이 뛰어나다.
		단 점	각 문서 데이터 베이스에서 용어들의 상호 관련성을 고려하지 못하고 있음 복합 용어들의 질의에 효과적이지 못한 경우가 많음.
신경망 학습방법	HOMIRS[d]	방 법	3층 구조의 다층 신경망을 구성하여 이를 바탕으로 임의의 질의에 대한 각 문서 데이터 베이스의 관련도를 계산하는 방법을 사용.
		장 점	복합용어로 구성된 질의가 대부분. 용어들간의 상관관계에 크게 영향을 받는 논문검색과 같은 전문정보검색에 뛰어난 성능을 보임.
		단 점	효과적으로 분산되지 않은 문서 데이터 베이스 환경에서 막대한 훈련 비용이 요구됨.

주) 비고: [a] Salton, 1971, [b] Gravano, Garcia-Molina, 1995, [c] Howe, Dreilinger, 1997, [d] 최용석, 2000.

도출하였다. 이 문제에 있어, 가장 단순한 방법은 사용자가 직접 각 문서 데이터베이스를 열람하는 방법이 있다. 널리 알려진 Prospero File System[Neuman, 1992], Gopher[Michael 등, 1992] 등이 이러한 방법을 사용하였다. 이 방법은 사용자가 원하는 정보를 세밀히 검색하는데 상당히 우수하나, 문서 데이터베이스가 증가할수록 필요 없는 문서 데이터베이스를 열람하는 소모비용이 증가하게 된다. 다른 방법은 다중플랫폼의 각 문서 데이터베이스에 대한 정보를 가지고있는 메타 데이터베이스를 구축하는 것이다. 따라서, 사용자는 그 메타 데이터베이스에 질의함으로써 열람하여야 할 문서 데이터베이스를 선택하여 열람하는 방식을 취한다. 예를 들면, WAIS[Kahle, Medlar, 1991]는 각 문서 데이터베이스를 요약하는 메타 데이터베이스로서 디렉토리 서비스[18]를 사용하였으며, ALIWEB[19] 시스템은 사용자가 생성한 각 문서 데이터베이스의 전체 내용에 대해 간략한 요약을 유지하는 방법을 사용한다. 이러한 방법의 유용성은 메타 정보를 얼마나 효과적으로 구축하는가에 달려있다. 다시 말해, 각 문서 데이터베이스의 모든 내용을 간략하면서도 완전하게 요약하는 것이 문제이며, 현실적으로 상당한 어려움을 안고 있다. 데이터베이스 선택문제 해결을 위한 선행 연구들을 정리하면 <표 1>과 같다.

18) Directory Of Service.
19) http://aliweb.emnet.co.uk/searchform.html.

18) Directory Of Service.
19) http://aliweb.emnet.co.uk/searchform.html.

2.2.4 문서 분류 및 카테고리

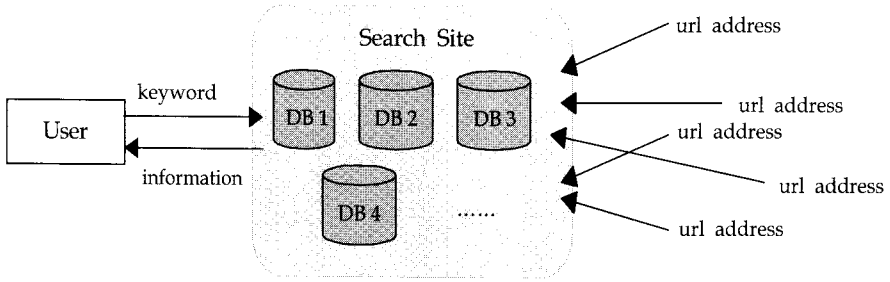
문서의 분류란 복수의 분류 카테고리를 정해 놓고 문서의 내용에 따라 하나 또는 그 이상의 카테고리를 문서에 지정함으로써 문서를 집단화하는 작업이다. 이러한 문서 분류는 과거에 대부분 수작업으로 이루어져왔으나, 온라인 문서의 양이 점차 많아지고 그 종류가 다양해지면서 문서 분류의 자동화에 대한 필요성이 널리 인식되었다[Blosseville 등, 1992].

일반적으로 문서를 자동 분류하는 방법에는 문서 내에 나타나는 단어의 빈도를 이용하여 분류 카테고리를 찾는 통계적 분류방법과 인간 전문가가 행하는 것처럼 문서의 내용을 기반으로 하는 분류규칙에 따라 분류를 수행하는 지식기반 분류방법이 있다. 현재는 사용자의 판단에 기초한 검색엔진으로 발전하여, 좀더 자율적인 오픈 소스카테고리와 사용자가 북마크한 자료를 활용한 검색서비스로 나타났다. 이러한 오픈소스카테고리를 활용한 검색서비스로는 넷스케이프사에서 운영하는 DMOZ[20]가 대표적이다. 이는 네티즌들이 직접 참여하여 자신의 관심 있는 분야에 대한 추천사이트를 등록하는 방식으로 운영하는 것이다. 또한, 검색의 질을 높이기 위해 각 카테고리별로 편집자 그룹을 모집하여 커뮤니티 방식으로 운영하고 있다. 현재, 넷스케이프의 Netcenter[21], Lycos, Altavista[22], Hotbot[23], AOL search[24] 등 약 90여개 사이트에서 직간접적으로 활용하고 있다. 북마크를 활용한 검색엔진에는 Clip2.com[25]와 Qbar[26]가 있으며 국내에는 북마크[27]와 서퍼[28] 등이 있다. 이러한 사이트들은 네티즌들의 북마크를 공유함으로써 이들을 취합하여 보편적인 카테고리 분류의 서비스를 제공하고 있다.

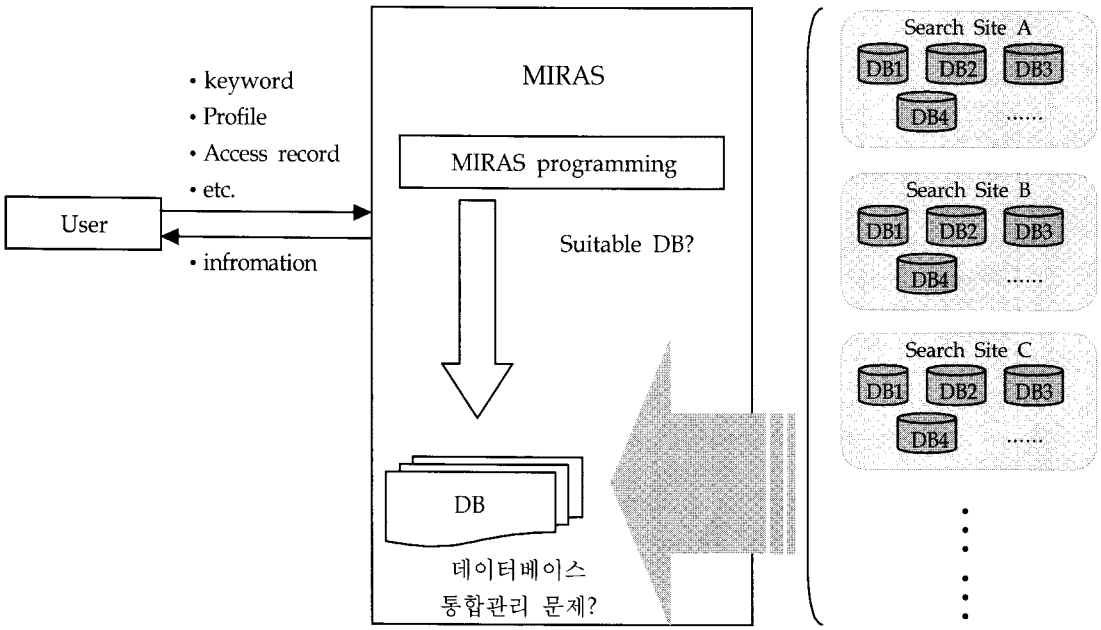
2.3 MIRAS의 정의 및 구조

정보검색 시스템에서 중요한 문제는 웹 환경하에서 필요한 정보를 사용자에게 정확하고 신속히 제공하는 기능이라 할 수 있다. 이러한 검색 에이전트가 정보를 검색하는 방법은 기존의 검색엔진을 활용하여 자료를 찾거나, 특정 키워드를 중심으로 직접 찾아 나가는 두 가지 방법이 있다. 전자의 경우는 여러 개의 검색엔진에서 나온 검색결과를 바탕으로 재정리하여 정보의 품질을 향상시킨다. 이 같은 방식을 MIRAS라 할 수 있다. 현재 이러한 서비스를 제공하는 대표적인 곳은 코페르닉[29]과 웹페럿[30]이 있으며 국내에는 와카노[31]가 있다. 모두 기존의 검색엔진 결과를 재구성하여 보여주는 메타검색서비스를 제공하고 있다. 또한, 이러한 데이터베이스는 각각 대량의 레코드를 지니고 있다. 가장 대표적으로는 각 검색사이트가 가지고 있는 URL 정보 등이 있다. 이들은 나름대로의 방식을 채택하여 데이터베이스를 분류하여 체계적으로 관리하고 있다. <그림 2>는 검색사이트의 데이터베이스 구조를 나타낸다. 각 검색 사이트의 데이터베이스는 단일 혹은 다중플랫폼의 형태로 존재한다. 이 가운데서 사용자가 찾고자 하는 문서 또는 사이트들에 적절한 데이터베이스를 선택하여 열람하는 것은 결코 쉬운 작업은 아니다. 특히, 메타검색서비스는 여러 검색 사이트들에 접근하여 사용자의 질의에 부합하는 자료를 가져옴으로써 검색질의에 보다 충실한 결과를 만드는 작업을 한다. 이러한 작업은 각 사이트에 존재하는 여러

20) http://dmoz.org/.
21) http://home.netscape.com/.
22) http://www.altavista.com/.
23) http://www.hotbot.com/.
24) http://www.AOL.com.
25) http://www.clip2.com/.
26) http://developer.ilikeq.com/qbar/.
27) http://search.bookmark.co.kr/.
28) http://www.surfer.co.kr/.
29) http://www.copernic.com.
30) http://www.zdnet.com/ferret/index.html.
31) http://www.wakano.co.kr/.



<그림 2> 검색 사이트의 데이터베이스 구조



<그림 3> MIRAS의 구조

개의 데이터베이스를 체계적으로 재구조화하는 것을 필요로 하며 효율적으로 접근하는 것을 요구한다. 이것이 MIRAS의 중요한 부분이라 할 수 있다.

<그림 3>은 이러한 MIRAS의 구조를 나타냈다. MIRAS는 사용자에게 일반 검색사이트 보다 나은 품질의 정보 및 서비스를 제공한다. 이를 위해서, 일반 검색 시스템보다 사용자에게 더 많은 정보를 요구하기도 한다. 가장 기본적인 관련 키워드정보를 비롯하여 사용자의 프로파일[32], 어세스레코드[33] 등을 사용자로부터 사전(事前)에 획득하거나 요구한다. 그러므로, MIRAS는 각 검색 사이트들이 제시하는 정보를 수집하여 자체 프로그래밍[34]으로 가공함으로써, 보다 개선된 품질의 정보를 사용자에게 제공한다. 여기서 시스템 자원 및 시간의 소모는 불가피하게 된다. 따라서, 이러한 소모를 최소화하는 동시에 적절한 정보를 사용자에게 제공하는 것이 MIRAS의 궁극적인 목표라 할 수 있다.

32) Profile.

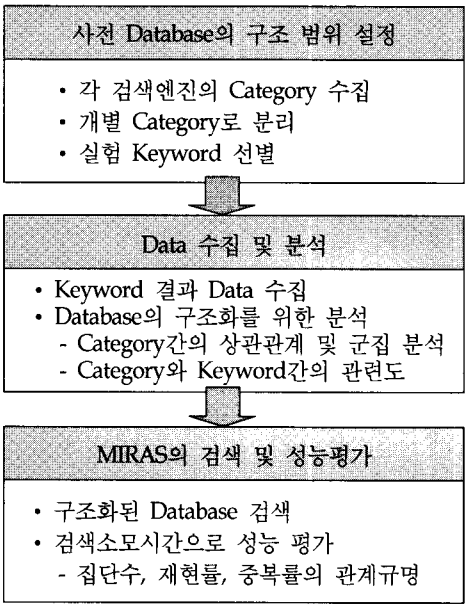
33) Access record.

34) Programming.

Ⅲ. MIRAS의 실험설계 및 데이터베이스 구조 분석

3.1 MIRAS 실험설계

본 연구에서 제시하는 MIRAS의 설계 구성도는 <그림 4>와 같이 세 부분으로 나눠진다.



<그림 4> MIRAS의 설계 구성도

첫째, 사전(事前) 카테고리의 구조 범위 설정이다. 이는 MIRAS가 응용할 기존의 각 검색 엔진들의 카테고리 구조를 사전에 파악하여, 그들 간의 관계를 규명하기 위한 작업에 해당된다. 여기서는 각 엔진들이 제시하는 카테고리를 조사하여 중복을 제외한 모든 카테고리를 개별 키워드 단위로 분리하는 작업을 수행한다.

둘째, 데이터의 수집 및 분석이다. 이 부분은 앞서 파악된 개별 카테고리를 기준으로 각 검색엔진에서 키워드ⓐ를 입력할 때 출현되는 데이터[35]를 수집한다. 수집된 데이터의 형태는 <그림 5>와 같다.

	CategoryA	Category B	CategoryC	CategoryD	CategoryE	...
Keyword A	32	18	32	2	2	-
Keyword B	28	10	25	3	1	-
Keyword C	10	32	5	18	17	-
...	-	-	-	-	-	-

<그림 5> MIRAS의 데이터 수집형태

이렇게 수집된 데이터는 MIRAS가 지향하는 체계적인 데이터베이스의 재구조화를 위한 분석에 이용된다. 분석의 핵심은 크게 두 가지로, 먼저 각 개별 카테고리간의 관계를 분석하여 재집단화 한다. 다음으로, 재집단화 된 카테고리를 기준으로 기존의 키워드에서 제시되었던 결과데이터를 재정리한다. 본 연구는 카테고리간의 상관관계분석으로 관련성을 규명할 수 있으며, 재집단화 및 개별 카테고리에서 나타나는 키워드ⓐ결과데이터의 수를 이용하여 관련도를 산출할 수 있다.

셋째, 이러한 환경에서 MIRAS의 검색 및 성능 평가이다. 이 부분은 앞서 분석한 결과를 토대로 검색을 수행하고 그 성능을 평가한다. 즉, 각 카테고리를 재집단화하여 다중플랫폼의 데이터베이스로 재구조화 하고, 키워드와 이들간의 관련도를 바탕으로 학습된 데이터베이스를 참조함으로써 검색을 수행한다. 이때의 시스템 소모시간, 결과의 수에 따른 재현률 등을 측정한다. 또한, 데이터베이스를 재구조화하는 과정에서는 데이터의 중복(Overlap)이 나타난다. 따라서, 최종적으로 재현률, 시스템 소모시간, 중복 등을 고려하여 MIRAS의 성능을 평가한다.

3.2 MIRAS의 성능 평가

MIRAS에서 재현률을 높이는 것은 MIRAS의 정보품질을 향상시키지만 시스템의 검색소모시

35) 이하, 키워드ⓐ결과데이터.

간의 손실은 불가피하게 된다. 또한 시스템의 검색소모시간은 MIRAS의 품질을 저하시키는 요인이다. 다시 말해, 정보검색환경에서 MIRAS의 재현률과 시스템의 소모시간 사이에는 배반관계(Trade-off)가 존재한다. 또한, 몇 개의 카테고리로 데이터베이스를 재구조화하는 과정에서 데이터의 중복은 불가피하게 나타난다. 데이터 중복은 시스템 성능을 저하시킨다. 본 연구는 이러한 MIRAS의 재현률과 시스템의 검색소모시간, 데이터 중복을 정의하고 고찰하고자 한다. 앞으로 다루게 될 MIRAS의 재현률, 시스템의 검색소모시간, 데이터 중복을 정의하면 다음과 같다.

MIRAS의 재현률(Recall Ability)은 사용자가 제시한 키워드로부터 각 개별 검색엔진에서 제시하는 관련 정보의 수[36]를 전체로 볼 때, 구조화된 데이터베이스를 가지고 있는 MIRAS에서 제시하는 정보의 총 수[37]를 비율로 나타낸다. 이는 식 (2)과 같다.

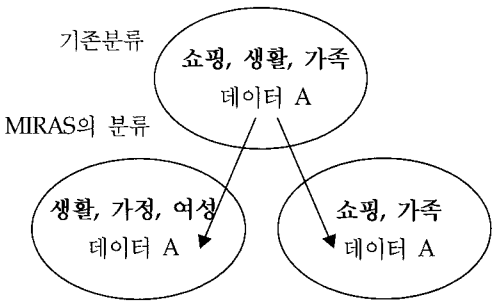
$$\text{MIRAS 재현율} \equiv \frac{\text{MIRAS가 제시하는 키워드(a)결과데이터 총 수}}{\text{키워드(a)결과데이터 총수}} \quad (2)$$

검색소모시간은 구조화된 MIRAS의 데이터베이스를 검색하는데 소요하는 시간을 의미한다. 실제로 검색순서를 결정하기 위한 각 키워드와 카테고리간의 관련도의 정렬부분 및 기타 연산은 각 집단마다 공통적인 소모시간이며, 본 연구의 초점에 벗어남으로 제외하였다. 따라서, MIRAS에서 검색소모시간은 재현률에 따른 관련 데이터베이스 하나를 열람하는데 소모되는 시간과 한 개의 레코드를 읽는데 소모되는 시간을 측정하여 전체를 검색하는데 소모시간을 추정할 수 있다. 이는 식 (3)과 같이 나타낼 수 있다.

$$\text{총 검색소모시간} \equiv \text{Intersept} + \beta \times \text{DBi} + \gamma \times \text{Ri} + \varepsilon \quad (3)$$

β : 한 개의 데이터베이스 열람시간
DBi : 데이터베이스 열람수
γ : 한 개의 레코드 검색 시간
Ri : 레코드의 수
ε : 교란항

데이터 중복은 구조화 과정에서 일어나게 된다. <그림 6>은 데이터 중복(overlap)의 예를 보여 준다.



<그림 6> 데이터 중복 예

즉, 기존의 분류 체계에서 한 개의 집단이 재구조화 과정에서 두개의 다른 집단으로 분류됨으로써 같은 데이터가 두개의 집단에 소속되는 현상이다. 이러한 중복은 식 (4)에 의해 산출된다.

$$\text{중복(Over lap)} = \text{집단(a)의 전체레코드 수} - \text{단일집단의 전체 레코드 수} \quad (4)$$

3.3 사전 카테고리 구조

각 검색엔진은 다양한 카테고리들에 따라 계층적으로 구성된 검색카테고리들을 가지고 있다. 검색카테고리들은 주어진 질의에 대해 특정한 카테고리의 관련 정보를 요약하여 제공함으로서, 각각 하나의 데이터베이스와 같이 동작한

36) 이하, 키워드(a)결과데이터 총 수.
37) 이하, MIRAS가 제시하는 키워드(a)결과데이터의 총 수.

라이코스(12)	네이버(14)	심마니(14)	야후(14)	엠파스(14)	한미르(12)
건강과 의학	건강과 의학	건강과 의학	건강과 의학	건강과 의학	건강과 의학
비즈니스경제	비즈니스경제	비즈니스경제	비즈니스경제	경제기업재테크	비즈니스경제
뉴스와 미디어	뉴스와 미디어	뉴스와 미디어	뉴스와 미디어	뉴스와 미디어	뉴스와 미디어
엔터테인먼트 예술	엔터테인먼트 예술	엔터테인먼트	엔터테인먼트		엔터테인먼트 예술
		예술	예술과 인문	문화 예술 종교	
사회 문화	사회 문화	사회생활	사회와 문화	정치 사회 법	사회 문화
가정과 생활	가정 여성				생활 가정
쇼핑	쇼핑	쇼핑		쇼핑 생활가족	
컴퓨터 인터넷	컴퓨터 인터넷	컴퓨터, 인터넷	컴퓨터, 인터넷	컴퓨터 인터넷	컴퓨터 인터넷
교육 참고자료	교육 참고자료	학문 참고자료	참고자료	사전 참고자료	교육 참고자료
		교육 학습	교육	학교 학원 교육	
				연예 오락 취미	
	게임	게임			
학문과학	학문과학		자연과학		과학 학문
			사회과학		
레크레이션 스포츠	스포츠	취미 스포츠	레크레이션 스포츠	여행 레저 스포츠	여행 레저 스포츠
	레크레이션				
	지역정보	지역정보	지역정보	지방 지역 세계	지역정보
		정치 행정	정부		
				사람찾기 개인홈페이지	
				학문학술단체	
취업정보					취업

<그림 7> 각 검색엔진들의 상위 검색 카테고리 현황

다. 본 연구에 사용되는 6개 검색엔진[38]의 상위 검색카테고리의 현황은 <그림 7>과 같다. 즉, 각 검색엔진은 조금 상이한 구조로 다중플랫폼의 데이터베이스를 가지고 있으며, 이를 관리한다. 라이코스와 한미르는 12개의 상위검색카테고리를 가지고 있으며, 네이버와 심마니, 야후, 엠파스는 14개의 상위검색카테고리를 가지고 있다. 상위검색카테고리의 구조를 보면 "*건강과 의학*", "*뉴스와 미디어*" 등 동일한 혹은 유사한 명칭의 카테고리를 가지거나, 엠파스의 "*사람찾기, 개인홈페이지*"와 같이 다른 구조로 나타나기도 한다.

MIRAS는 이러한 명확하지 않은 분류환경에서 기존의 검색엔진이 가지고 있는 데이터베이스를 통합하여 재구조화하기 위해 상위검색카테고리를 키워드 단위로 분리한다. 키워드 단위로 분리한 결과는 <그림 8>과 같다. 이로서, 최소 단일플랫폼에서 각각의 카테고리 키워드로 구성된 51개의 다중플랫폼으로 데이터베이스를 통합 구조화 할 수 있다.

38) 네이버, 라이코스, 심마니, 야후, 엠파스, 한미르.

카테고리 키워드
건강(b1), 의학(b2), 병원(b3), 비즈니스(b4), 경제(b5), 기업(b6), 재태크(b7), 뉴스(b8), 미디어(b9), 엔터테인먼트(b10), 예술(b11), 인문(b12), 문화(b13), 종교(b14), 사회(b15), 생활(b16), 정치(b17), 법(b18), 가정(b19), 여성(b20), 쇼핑(b21), 가족(b22), 컴퓨터(b23), 인터넷(b24), 교육(b25), 참고자료(b26), 학문(b27), 사전(b28), 학습(b29), 학교(b30), 학원(b31), 연예(b32), 오락(b33), 취미(b34), 게임(b35), 과학(b36), 자연(b37), 레크레이션(b38), 스포츠(b39), 여행(b40), 레저(b41), 지역정보(b42), 지방(b43), 세계(b44), 행정(b45), 정부(b46), 사람찾기(b47), 개인홈페이지(b48), 학술(b49), 단체(b50), 취업(b51)

<그림 8> 상위 검색 카테고리의 키워드 분리

순위	라이코스	야 후	네이버	한미르	심마니	엠파스	드림위즈
1	만화	벅스뮤직	디아블로	게임	와레즈	벅스뮤직	야후
2	엽기	다음	리니지	엽기	엽기	숙박업	엽기
3	와레즈	세이클럽	뮤	디아블로	게임	엽기	다음
4	디아블로 2	리니지	엽기	세이클럽	청소년보호위원회	성인	벅스뮤직
5	야후	다모임	청소년보호위원회	졸라맨	디아블로	채팅	뮤
6	벅스뮤직	엽기	포트리스	뮤	네오지오	요식업	게임
7	다음	넷마블	채팅	그림	리니지	유머	세이클럽
48	아이러브스쿨		케릭터			스타크래프트	세이
49	포트리스		유머			교체	마시마로
50	음악		이력서			지도	다음

자료) 2001년 10월.

<그림 9> 인기검색어 순위표

3.4. 실험검색 키워드

본 연구에서는 각각 유사한 카테고리들간에 집단화 하여 효율적인 MIRAS 설계를 실험해 보기 위하여 실험 데이터를 수집하였다. 먼저, 인기검색어에서 실험 키워드 선별은 <그림 9>와 같이 라이코스, 네이버, 엠파스, 드림위즈에서 제시하는 인기검색어 50위의 순위 자료와 야후, 한미르 심마니의 30위의 순위 자료에서 선별하였다.

라이코스의 인기검색어인 "*와레즈(3위)*"나 "*디아블로2(4위)*", "*야후(5위)*" 등은 특정 사이트를 나타내는 검색어로서 데이터 수집 목적에 부합하지 않으므로 인기키워드선별에서 제외하였다. 그러한 기준에 따라 선별된 인기키워드는 <그림 10>과 같다. 따라서, 실험 데이터 수집에는 <그림 8>의 카테고리 키워드 51개와 인기검색어에서 60개를 선별하여 사용하였다.

실험 분석에 사용된 데이터는 한미르, 야후, 네이버, 라이코스, 심마니, 엠파스의 검색 엔진에서 추출하였다. 앞에서 선별된 카테고리 키워드와 인기키워드(전체 111개)에 따른 각 검색 엔진에서 제시하는 상위 결과 20개씩의 사이트주소, 제목, 웹디렉토리, 디렉토리 주소, 엔진명, 설

인 기 키 워 드
만화, 엽기, 뮤직비디오, 무료, 영화, mp3, 노래방, 채팅, 자료실, 지도, 신문, 개인방송, 아르바이트, 자바, 십자수, 핸드폰, 음악 배경화면, 퀴즈, 성인방송, 중고차, 운세, 태그, 부동산, 음악감상, 독후감, 동영상, 환경오염, 하늘, 캐릭터, 유머, 이력서, 그림, 우편번호, 과외, 숙박업, 요식업, 성인, 협회, 맥주, 소주, 에너지, 메일, 음료, 선물, 카메라, 공포, 동물, 자동차, 건설, 팬티, 연극, 교체, 사진, 코리아, 검색, 다운, 딴따라, www, 세이

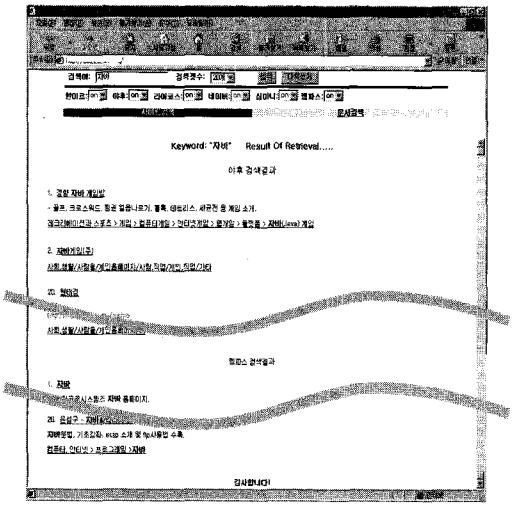
<그림 10> 인기검색어의 선별

명을 수집하여 분석에 사용하였다.

Ⅳ. 데이터 수집 및 분석

4.1 데이터 수집 환경

데이터 수집 환경은 자바 웹 서버(Java Web Server 2.0)에서 자바 서블릿(Java Servlet)으로 구현하였으며, 어세스(Microsoft Access)를 이용하여 데이터베이스를 구축하였다. <그림 11>은 MIRAS에서 키워드에 따른 데이터를 수집하는 인터페이스를 보여준다.



<그림 11> MIRAS의 데이터 수집 인터페이스

이러한 인터페이스를 이용하여 데이터를 수집한 결과는 <그림 12>에서와 같이 단일플랫폼의 형태로 데이터베이스에 저장된다. 이러한 단일플랫폼의 형태의 표본 데이터를 이용하여 카테고리 간의 관계를 분석하고, 키워드와 카테고리의 관계를 분석한다.

<그림 12> 데이터베이스 파일

4.2 카테고리간의 관련도

본 연구에서 카테고리간의 관련도를 구하기 위하여 <그림 13>의 데이터를 가지고 상관분석을 실시하였다. 분석에는 SAS Ver.8.0[39]을 사용하였다.

카테고리간의 표본상관계수를 나타내면 <그림 14>와 같다. 여기서, 유사성을 바탕으로 다수의 카테고리 키워드를 몇 개의 동질적인 새로운 카테고리로 구분 할 수 있는 군집분석을 실시하였다. 군집분석은 SAS 의 PROC VARCLUS[40]를 사용하였다. 결과적으로 군집분석에서 설명률을 참조함으로써 다중플랫폼의 데이터베이스의 재구조의 대안을 설정한다.

군집분석의 설명력은 <그림 15>와 같다. 그림에서 나타나듯이 13개 새로운 카테고리로 분류될 때부터 전체에 대한 설명력의 증가폭이 완만

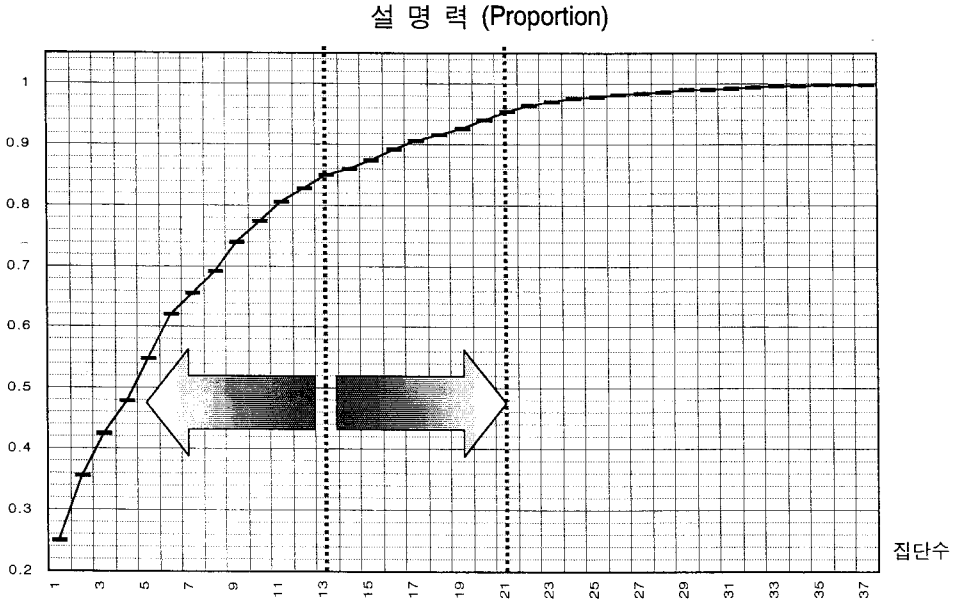
39) Statistical Analysis System.
40) 상관관계행렬을 이용한 계보적 군집방법.

Keyword	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10	b11	b12	b13	b14
건 강	32	32	8	17	17	0	0	0	0	1	1	0	0	0
의 학	32	32	9	5	5	0	0	0	0	1	1	0	0	0
병 원	23	23	0	7	17	10	10	0	0	0	0	0	0	0
비즈니스	0	0	0	30	32	2	2	1	0	8	7	0	0	0

<그림 13> 각 키워드의 상위 카테고리의 출현 수

	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10
b1	-	1.00	0.88	-0.13	-0.12	-0.05	-0.05	-0.08	-0.07	-0.13
b2	-	-	0.88	-0.13	-0.12	-0.05	-0.05	-0.08	-0.07	-0.13
b3	-	-	-	-0.0.6	-0.08	-0.05	-0.13	-0.17	-0.06	-0.11
b4	-	-	-	-	0.98	0.69	0.69	-0.11	-0.11	-0.29
b5	-	-	-	-	-	0.81	0.81	-0.11	-0.13	-0.29
b6	-	-	-	-	-	-	1.00	-0.10	-0.14	-0.23
b7	-	-	-	-	-	-	-	-0.10	-0.14	-0.23
b8	-	-	-	-	-	-	-	-	0.97	0.01
b9	-	-	-	-	-	-	-	-	-	0.01

<그림 14> 카테고리간의 표본상관계수



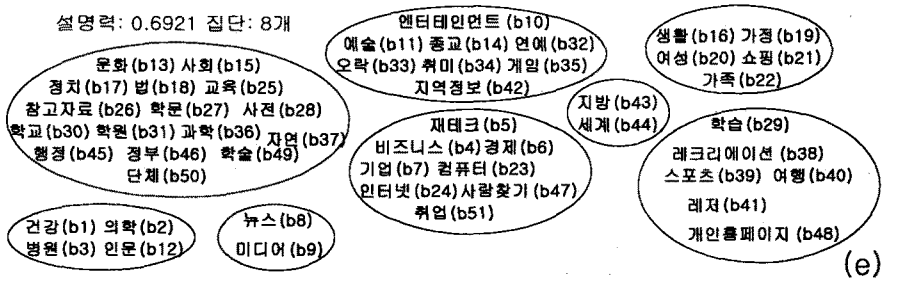
<그림 15> 카테고리 키워드의 군집분석의 설명력

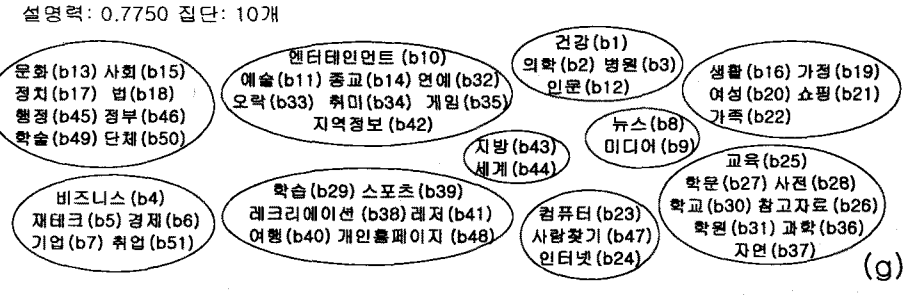
※ 출력범위 20개. SAS를 통한 군집분석(Varclus)

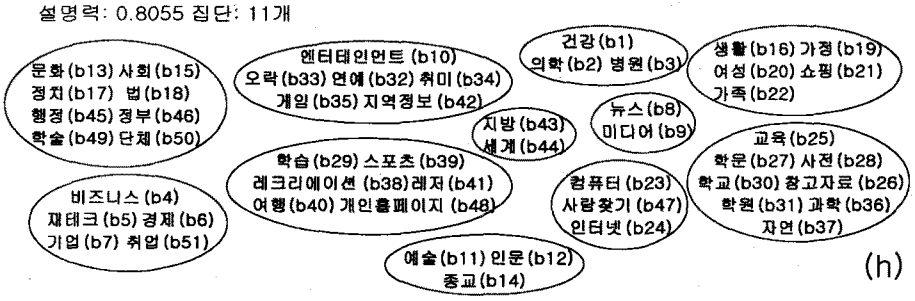
(a) 설명력: 0.4238 집단: 3개

⋮

(d) 설명력: 0.5480 집단: 7개







(i) 설명력: 0.8281 집단: 12개
(j) 설명력: 0.8599 집단: 14개
(k) 설명력: 0.9060 집단: 17개
(l) 설명력: 0.9524 집단: 21개

⋮

<그림 16> 카테고리 키워드의 PROC VARCLUS 분석결과

keyword	건 강	의 학	병 원	비즈니스	경 제	기 업	제테크
카테고리 1	0.01	0.01	0.01	0.04	0.09	0.05	0.07
카테고리 2	0.00	0.00	0.00	0.00	0.01	0.00	0.00
카테고리 3	0.00	0.00	0.00	0.00	0.00	0.00	0.00
카테고리 4	0.00	0.00	0.00	0.00	0.00	0.00	0.00
카테고리 5	0.01	0.01	0.00	0.10	0.00	0.00	0.00
카테고리 6	0.26	0.11	0.30	0.61	0.34	0.82	0.72
카테고리 7	0.62	0.57	0.41	0.00	0.00	0.00	0.00
카테고리 8	0.05	0.00	0.01	0.00	0.01	0.00	0.07
카테고리 9	0.01	0.01	0.00	0.05	0.00	0.00	0.04
카테고리 10	0.00	0.00	0.00	0.00	0.02	0.02	0.00
카테고리 11	0.01	0.03	0.02	0.02	0.00	0.02	0.00
카테고리 12	0.01	0.09	0.05	0.07	0.09	0.02	0.04
카테고리 13	0.00	0.11	0.02	0.00	0.16	0.01	0.00
카테고리 14	0.01	0.00	0.00	0.02	0.05	0.00	0.01
카테고리 15	0.00	0.01	0.00	0.00	0.17	0.00	0.00
카테고리 16	0.02	0.04	0.19	0.07	0.04	0.04	0.05
카테고리 17	0.00	0.00	0.00	0.00	0.01	0.01	0.00

.....

<그림 17> 키워드와 17개 새로운 카테고리간의 관련도

해짐을 알 수 있다. 또한 21집단은 95% 이상의 설명률을 가짐으로 최대 21집단의 분류까지 대안을 설정하였다. 이와 같이 군집분석을 응용함으로써 다중플랫폼 데이터베이스의 재구조에 대안을 설정하는 것을 목적으로 한다.

카테고리 키워드의 군집분석결과는 <그림 16>과 같다. 그림에서 보듯이 카테고리의 수가 증가할수록 전체를 표현하는 설명력이 높아짐을 알 수 있다. 51개의 개별 카테고리는 이러한 유사성을 바탕으로 (a)와 같이 전체에 대한 0.4238의 설명력을 가지는 3개 새로운 카테고리로도 나타낼 수 있으며, (l)와 같이 전체에 대한 0.9524의 설명력을 가지는 21개 새로운 카테고리로 대안을 설정할 수 있다.

MIRAS에서 적은 집단으로 재구조화하는 것은 검색에서 열람해야 할 데이터베이스의 대안의 수가 감소하여 속도를 향상시키거나, 시스템의 자원소모를 줄일 수 있다. 그러나, 동시에 한 개의 데이터베이스에 관련 없는 정보의 레코드 수가 증가함으로써 속도가 저하되거나, 시스템의 자원소모가 증가한다. 이러한 관계를 좀더 실증적인 분석을 통하여 고찰하려 한다. 결과로 효과적인 다중플랫폼 데이터베이스의 재구조화 방안을 도출한다.

4.3 키워드와 카테고리간의 관련도

키워드와 카테고리간의 관련도는 다음과 같이 산출한다. MIRAS에서 각 키워드로 검색하였을 때, 나타나는 카테고리의 출현 수의 데이터를 분석하여, 카테고리와 키워드와의 관련도를 식 (5)에 의해 산출한다.

$$\text{키워드(a)와 카테고리(A)의 관련도} \equiv \frac{\text{카테고리(A)에 소속되 있는 키워드(a)결과데이터수}}{\text{키워드(a)결과데이터 총 수}} \quad (5)$$

17개의 새로운 카테고리로 구조화할 경우 각 키워드와 카테고리간의 관련도는 식 (5)에 따라 <그림 17>과 같이 나타낼 수 있다.

이러한 형태로 새로운 카테고리분류 방식에서 각 키워드와의 관련도를 산출할 수 있다. 따라서, MIRAS의 검색에서는 이미 학습된 키워드를 입력 받았을 때, 관련도가 높은 카테고리순으로 검색작업이 이루어지게 된다. 식 (6)에 의하여 관련도의 합이 데이터의 재현률을 나타냄을 알 수 있다. 즉, MIRAS는 내부적으로 관련도 수준에 따라, 재현률을 조절할 수 있게 된다.

키워드(a)의 MIRAS 재현률 ≡

Σ 키워드(a)와 각 카테고리의 관련도 (6)

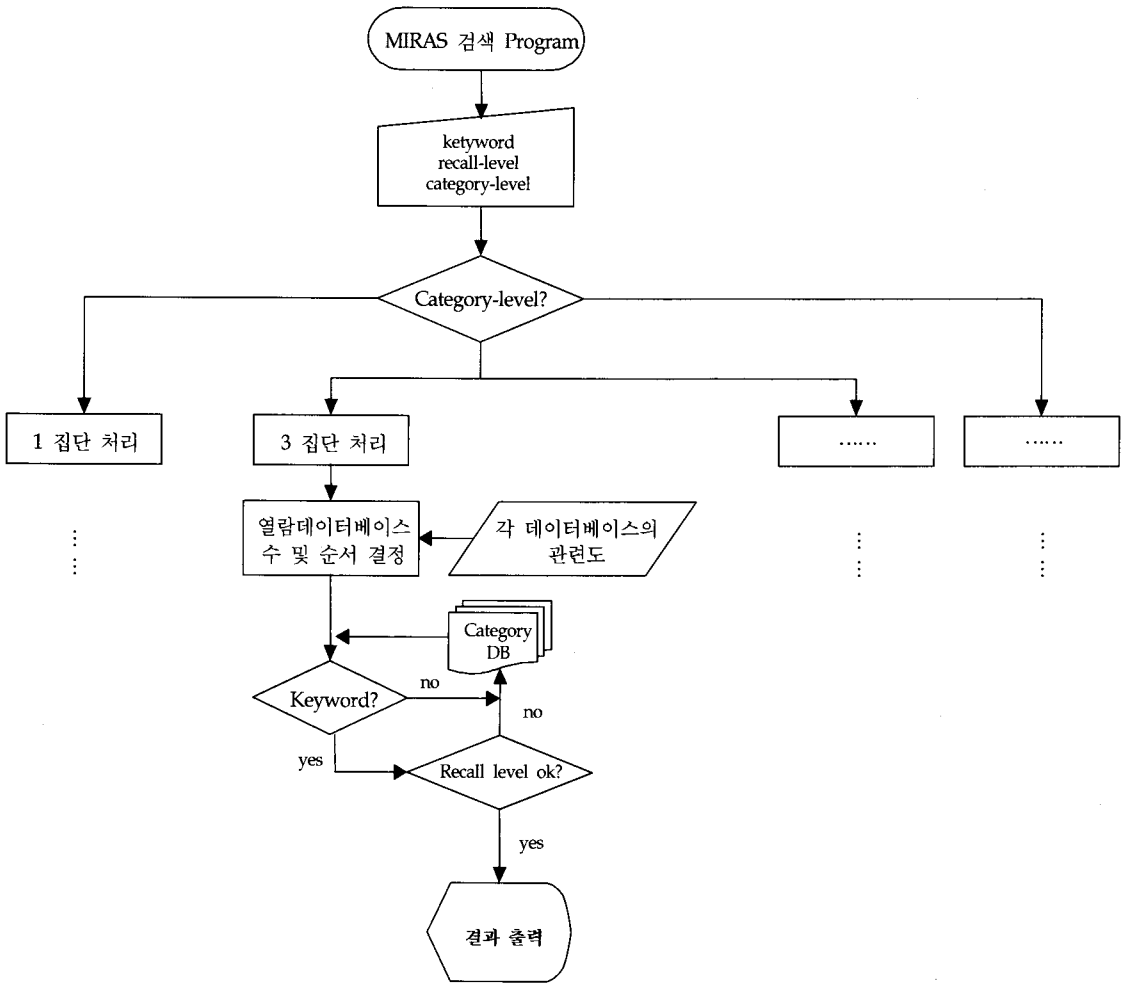
∵ MIRS가 제시하는 키워드(a)결과데이터총 수

≡ Σ각 카테고리에 소속되 있는 키워드(a)결과 데이터수

V. MIRAS의 성능 평가

5.1 MIRAS의 성능평가

앞장에서의 분석을 토대로 구조화된 데이터



<그림 18> 검색 프로그램 순서도

베이스를 가지는 MIRAS의 성능 평가를 위하여 검색을 수행하는 프로그램의 순서도를 나타내면 <그림 18>과 같다. 즉, 본 검색 프로그램은 재구조화된 데이터베이스를 가지는 MIRAS의 성능을 평가하기 위한 목적으로 설계하였다

최초 사용자가 키워드와 카테고리 분류수준, 재현률 수준을 입력하였을 때, 카테고리 분류수준에 맞는 다중플랫폼의 데이터베이스를 구성한다. 이후, 각 데이터베이스와 키워드간의 사전에 학습된 관련도를 참고하여 검색할 데이터베이스의 수와 순서를 결정하여 검색을 수행한다. 이때, 사전에 입력한 재현률 수준에 결과의 값이 도달하면 출력하게 된다.

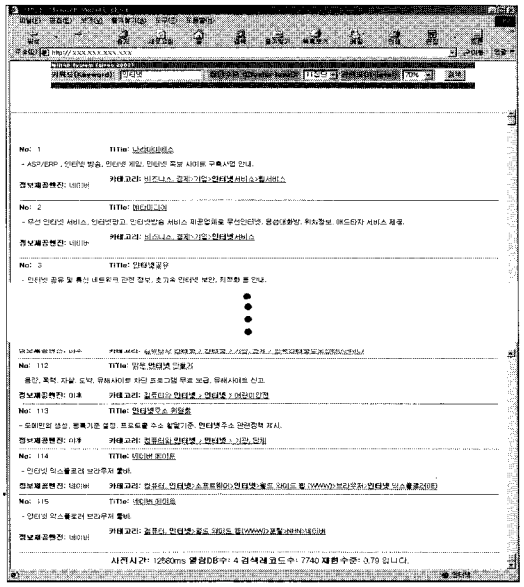
본 프로그램은 자바 서블릿으로 구현하였으며, 시스템 환경은 CPU: PentiumⅢ-450의 메모리: 256M에서 자바 웹 서버 2.0을 사용하였다. 자체 프로그램의 단계별 소모시간측정을 위하여 프로그램코드의 부분에 프로파일코드를 삽입함으로써 각 부분의 실행 시 소모시간을 측정하였다. 자바 프로그래밍에서 시간측정 코드는 <그림 19>와 같다. 여기서 측정단위는 ms[41]단위로 측정된다.

```
Long start = System.currentTimeMilis( );

//시간 측정을 할 연산들
… …
… …
Long time = System.currentTimeMilis( )-start;
```

<그림 19> 시간측정 코드

<그림 20>은 MIRAS에서 검색시에 사용되는 사용자 인터페이스를 나타낸다. 사용자가 초기화면에서 검색키워드와 재현률의 수준, 카테고리의 수준을 결정하여 입력하면 그에 따른 정보결과와 검색결과의 수, 데이터베이스의 열람 개수, 총검색레코드 수, 검색수행 시 소모된 시간[42]을 출력한다



<그림 20> 사용자 검색 인터페이스

5.2 MIRAS의 데이터 분석

앞 절의 사용자 검색 인터페이스로부터 <그림 21>과 같은 검색결과를 얻을 수 있다. 그림에서 (a-1), (a-2), (a-3), (a-4), (a-5), (a-6)은 검색키워드인 '건강'으로 각각 100%, 90%, 80%, 70%, 60%, 50%의 재현률 수준으로 검색을 수행한 결과이다.

또한, 검색결과에서 평균시간은 각 집단 별로 5번씩 수행한 결과 소모시간의 평균을 나타낸다. 이때 열람한 데이터베이스의 개수[43]와 검색된 레코드 수[44]를 각각 *'DB'*와 *'레코드수'*로 나타내었다.

실제 재현률을 단일플랫폼의 검색결과와 집단(a)의 검색결과의 비율로서 산출하였다. MIRAS

41) Millisecond(0.001초).

42) 이하, 검색소모시간.
43) 이하, 열람데이터베이스 수.
44) 이하, 검색레코드 수.

에서 각 키워드에 따른 실제 재현률은 표에서 나타나듯이 집단분류의 수가 높을수록 사용자가 원하는 수준에 가까이 도달한다. <그림 21>의 MIRAS의 검색결과를 정리하여 소모시간의 분석 데이터를 수집하면 <그림 22>와 같다.

이 데이터를 이용하여 각 변수간의 상관 관계

	1 집단	3 집단	5 집단	6 집단	9 집단	11 집단	14 집단	17 집단	21 집단
1회	8400	16210	19000	17580	19380	20630	18890	18560	20540
2회	8620	16150	18450	17410	18950	19990	18780	18240	19500
3회	8400	16090	18620	17520	19110	20380	19390	19000	20980
4회	8510	16100	18290	17420	18560	20540	18570	19390	19610
5회	8510	15820	18340	17190	18620	20110	19170	18350	19440
평균시간	8488	16074	18540	17424	18924	20330	18960	18708	20014
재 현 률	1	1	1	1	0.99	0.99	0.99	0.99	0.99
DB	1	3	5	5	7	8	9	8	9
레코드수	12061	13148	13001	12273	10947	10672	10036	9714	10598
검색결과	125	132	131	128	126	126	123	124	125

(a-1) 검색키워드 '건강'으로 100%의 재현률 수준으로 검색을 수행한 결과

⋮

(a-2) 검색키워드 '건강'으로 90%의 재현률 수준으로 검색을 수행한 결과
(a-3) 검색키워드 '건강'으로 70%의 재현률 수준으로 검색을 수행한 결과
(a-4) 검색키워드 '건강'으로 60%의 재현률 수준으로 검색을 수행한 결과

⋮

	1 집단	3 집단	5 집단	6 집단	9 집단	11 집단	14 집단	17 집단	21 집단
1회	-	5270	2800	2800	2360	2140	2310	2300	2190
2회	-	4950	2900	2800	2530	2250	2310	2200	2140
3회	-	5220	2800	2810	2250	2200	2580	2190	2140
4회	-	5110	2860	2910	2250	2200	2150	2200	2200
5회	-	5170	2750	2800	2310	2200	2200	2200	2200
평균시간	-	5144	2822	2824	2340	2208	2310	2218	2174
재 현 률	1	0.66	0.63	0.63	0.62	0.62	0.62	0.62	0.62
DB	1	1	1	1	1	1	2	1	1
레코드수	12061	1397	938	938	493	410	410	410	410
검색결과	125	86	82	82	78	78	78	78	78

(a-5) 검색키워드 '건강'으로 50%의 재현률 수준으로 검색을 수행한 결과

주) 비고: 평균시간의 단위는 ms.

<그림 21> MIRAS의 검색 결과

no	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
시간	8488	16074	18540	17424	18924	20330	18960	18708	20014	10496	9798	9424	8972	8116	9196	7438	7594	7132	6122
DB	1	3	5	5	7	8	9	8	9	2	3	3	3	3	4	3	2	2	2
레코드수	12061	13148	13001	12273	10947	10672	10036	9714	10598	8730	6639	3559	6114	4824	4824	4221	5747	5302	4012
집 단 수	1	3	5	6	9	11	14	17	21	3	5	6	9	11	14	21	5	9	11
검색결과	125	132	131	128	126	126	123	124	125	122	121	121	117	116	116	116	117	113	112

no	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
시간	6130	5954	12063	12211	20154	21254	19926	19018	20882	14092	1524	15864	14048	13350	12326	11910	11952	11150	10764
DB	3	2	4	5	7	8	9	8	10	3	4	5	5	6	5	5	4	4	5
레코드수	1012	3901	12063	12211	12401	12089	11454	10264	9661	11171	11399	9896	8269	7726	6748	6291	7649	6442	5902
집 단 수	14	21	5	6	9	11	14	17	21	5	6	9	11	14	17	21	9	11	14
결색결과	112	112	146	146	146	145	145	144	144	145	145	144	139	138	137	132	126	121	121

no	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53
시간	10446	10260	9246	10620	10476	9646	9008	8906	8780	6392	6018	6296	5842	4986	5004
DB	4	4	2	3	3	3	4	3	3	1	1	2	2	3	2
레코드수	5902	5791	6362	6590	7632	6425	5885	5885	5774	4418	4098	3140	2823	2823	2823
집 단 수	17	21	5	6	9	11	14	17	21	3	5	6	9	14	17
검색결과	121	121	101	101	125	120	120	120	120	83	82	99	81	81	81

<그림 22> 소모시간의 분석 데이터

분석을 실시한 결과는 <그림 23>과 같다.

	소 모 시 간	DB	레코드 수	집단수	검 색 결 과
소 모 시 간	1.0000				
DB	0.8883 < .0001	1.0000			
레코드 수	0.8396 < .0001	0.5842 < .0001	1.0000		
집단수	0.0759 0.5892	0.4066 0.0025	-0.2647 0.0555	1.0000	
겸 색 결 과	0.7224 < .0001	0.6041 < .0001	0.7673 < .0001	0.0555 0.6933	1.0000

<그림 23> 각 변수간의 상관관계

소모시간과 각각 열람데이터베이스 수, 검색레코드 수, 검색결과 간의 상관계수 값이 높게 나타난다. 그러나, 검색결과는 각각 검색레코드 수, 열람데이터베이스 수간의 상관계수 값도 높게 나타난다. 따라서, MIRAS에서 검색프로그램의 소모시간에 영향을 미치는 요인으로 데이터베이스를 열람하는데 소모되는 시간과 레코드를 읽는데 소모되는 시간을 고려한다. 즉, 총 소모시간을 종속변수로 하고 열람데이터베이스와 검색레코드 수를 독립변수로 하는 Model(1)을 설정할 수 있다.

Model(1):

총 소모시간 = Intercept + $\beta \times DBi$
+ $\gamma \times Ri + \varepsilon$

Intercept : 기타 연산과정
β : 한개의 데이터베이스 열람시간
DBi : 데이터베이스 열람수
γ : 한개의 레코드 검색시간
Ri : 레코드의 수
ε : 교란항

여기서 인터셉(Intercept)은 데이터베이스의 열람우선 순위를 결정하기위한 정렬 및 기타 연산과정의 값을 의미한다. <그림 22>의 데이터를

분산 분석표

Source	DF	Squares	Square	F Value	Pr > F
Model	2	1192764790	596382395	430.92	< .0001
Error	50	69198651	1383973		
Corrected Total	52	1261963441			

Root MSE: 1176.42383 R-Square: 0.9452
Dependent Mean: 11929 Adj R-Sq: 0.9430
Coeff Var: 9.86171

회귀계수 추정치

Variable	DF	Estimate	Error	T Value	Pr > |t|
Intercept	1	733.93645	432.72887	1.70	0.0961
DBi	1	1283.95663	86.73432	14.80	< .0001
Ri	1	0.75991	0.06371	11.93	< .0001

<그림 24> 다중회귀분석 결과

이용한 다중회귀분석으로 Model(1)을 검정하였다. 다중회귀분석결과는 <그림 24>와 같다. 각 추정치에 대한 값의 유의도를 살펴보면 각 모수에 대해 모두 수준에서 의미가 있다. 다중결정계수[45]는 0.9452로 다중회귀 모형이 약 94.52% 설명하고 있다.

따라서, 한 개의 데이터베이스를 열람하는 데 소모되는 시간과 한 개의 레코드를 읽는데 소모되는 시간을 Model(2)로 추정할 수 있다. 그러므로, Model(2)를 참조하여 나머지 검색에 소모되는 시간들을 추정할 수 있다.

Model(2):

$$\text{총 소모시간} = 733.93645 + 1283.95663 \times DBi + 0.75991 \times Ri + \varepsilon$$

DBi: 데이터베이스 열람수
Ri : 레코드의 수
ε : 교란항

즉, 검색시에 열람하는 데이터베이스의 개수와 검색 레코드의 개수로부터 본 연구의 검색 프로그램의 수행 시에 소모시간을 산출할 수 있다. 또한, 식 (7)에 의해 데이터베이스를 하나 열람하는데 소모시간과 1689.6167개의 레코드를 읽는데 소모시간과 대등함을 알 수 있다.

$$Ri = \text{데이터베이스 열람시간} \,/\, \text{레코드 검색시간} \quad (7)$$

Ri : 1개의 데이터베이스를 열람하는 소모시간과 대응하는 레코드의 수

따라서, 본 시스템 환경에서는 대략적으로 데이터베이스를 하나 더 열람함으로써 1690개 이상의 검색 대상의 레코드를 감소시킴은 검색소모시간을 단축시키는 결과를 나타낸다.

<그림 25>는 각 키워드와 각 집단별 열람데이터베이스 수와 검색레코드 수를 측정하고, Model(2)에 대입하여 소모시간을 산출한 데이터이다.

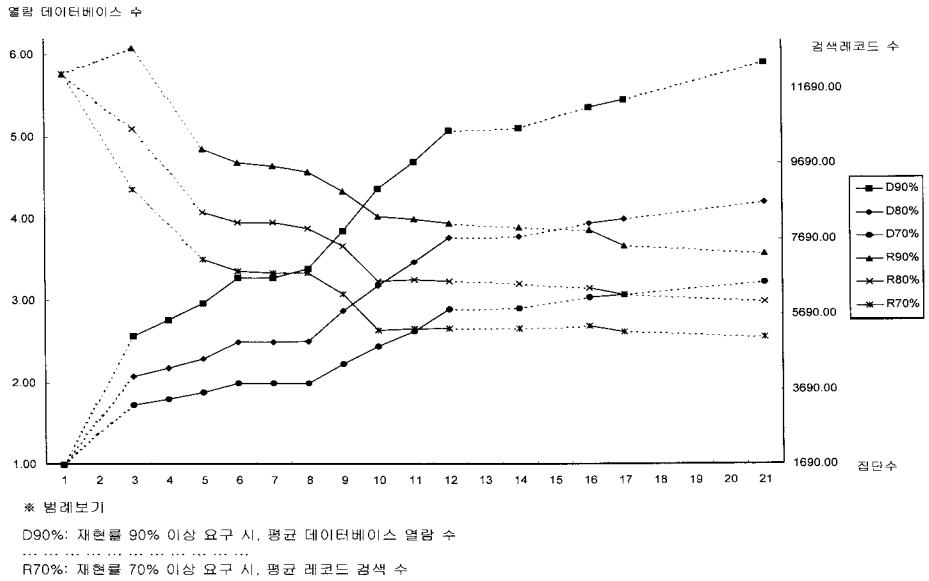
5.3 MIRAS의 성능 분석

이러한 결과를 토대로 실험검색키워드(111개)에 대한 재현률 90%이상, 80%이상, 70%이상 요구 시, 각 집단별 평균열람데이터베이스의 수 및 평균검색레코드 수는 <그림 26>과 같다.

MIRAS의 검색 수행시, 재현률의 요구 수준이 낮을수록 검색레코드 수와 열람데이터베이스 수

45) R^2

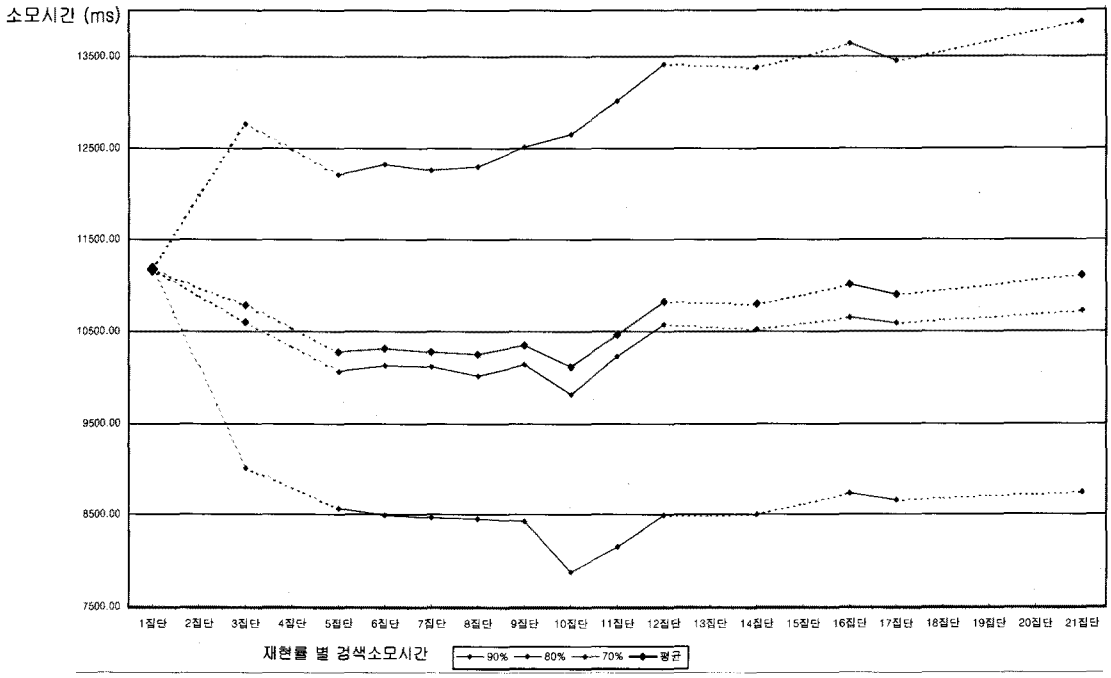
keyword	3집단	열람 DB	검색레코드	소모시간	5집단	열람DB	검색레코드	소모시간
건강		2	8730	9935.86		3	6639	9630.8
의학		3	13148	14577.10		3	9845	12067.1
병원		3	13148	14577.10		4	12109	15071.5
비즈니스		3	13148	14577.10		3	11171	13074.7
경제		3	9851	14577.10		3	11171	13074.7
기업		2	9851	10787.72		2	8907	10070.3
재테크		2	13148	10787.72		3	9799	12032.1
뉴스		3	13148	14577.10		3	9845	12067.1
미디어		3		14577.10		3	11171	13074.7



<그림 26> 평균 검색레코드 수 및 평균 열람데이터베이스 수

는 낮다. 또한, 각 집단별 평균열람데이터베이스 수는 집단수가 많을수록 증가하고, 각 집단별 평균검색레코드 수는 집단수가 많을수록 감소한다. 여기서, 그 증감의 폭을 자세히 살펴보면 평균검색레코드 수는 10집단에서 급감하다가 11집단에서 완만해진다. 평균열람데이터베이스 수는 12집단에서 완만히 증가함을 알 수 있다. 또한 17집단에서도 평균검색레코드 수는 비교적 급감하나, 그 수준이 10집단에 미치지 못하며, 평균열람데이터베이스 수는 높은 수준에 있다.

<그림 27>은 이러한 관계를 평균검색소모시간 분석으로서 명확히 나타낸다. 재현률 90%이상 요구 시에서는 단일플랫폼의 데이터베이스가 가장 낮은 검색소모시간을 가지지만, 재현률 80%와 70%이상 요구 시에서 다중플랫폼 데이터베이스는 관련 없는 레코드의 검색을 효과적으로 제외시킴으로써 평균레코드검색소모시간을 감소시킨다. 그러나 집단수가 증가할수록 평균열람데이터베이스의 수는 증가함으로써 그에 따른 검색소모시간은 증가한다.



재현율	1집단	2집단	3집단	4집단	5집단	6집단	7집단	8집단	9집단	10집단	11집단	12집단	14집단	16집단	17집단	21집단
100%	11176.33	13744.24	16312.15	17596.11	18880.07	20164.02	21447.98	22731.94	24015.89	25299.85	27867.76	30435.68	31719.63	36855.46		
90%	11176.33	12751.88	12202.31	12316.79	12259.01	12290.06	12500.97	12640.70	13014.03	13414.99	13369.30	13638.96	13445.01	13877.13		
80%	11176.33	10601.99	10057.81	10121.03	10110.50	10010.01	10129.11	9813.98	10217.06	10569.46	10528.25	10650.60	10587.21	10719.88		
70%	11176.33	9009.47	8571.10	8486.36	8464.82	8453.90	8427.04	7882.49	8151.40	8487.70	8504.45	8733.11	8659.21	8737.72		
평균	11176.33	10787.78	10277.07	10308.06	10278.11	10251.32	10352.37	10112.39	10460.83	10824.05	10800.67	11007.56	10897.14	11111.58		

<그림 27> 평균 검색소모시간 분석

이러한 결과로 각 집단별 검색소모시간의 평균을 살펴보면 전체 카테고리에서 10집단의 분류가 이상적임을 알 수 있다. 또한, 11집단 이상에서는 검색소모시간이 증가하다가 17집단에서 약간의 감소를 보이고 다시 증가함을 알 수 있다. 그러나, 검색소모시간으로 볼 때 17집단은 10집단에 못 미치는 감소를 나타낸다.

<그림 28>은 각 집단별 데이터의 분포 수 및 중복 수를 나타낸다. 실제 단일집단에 레코드 수는 12052개를 가지고 있으나 3집단으로 구조화 되었을 때, 레코드의 수는 13148개로 1096개의 데이터가 중복 생성됐음을 알 수 있다. 이러한 중복은 앞서 설명한 바와 같이 데이터베이스의 집단화 과정에서 불가피하게 생성되는 데이터이다.

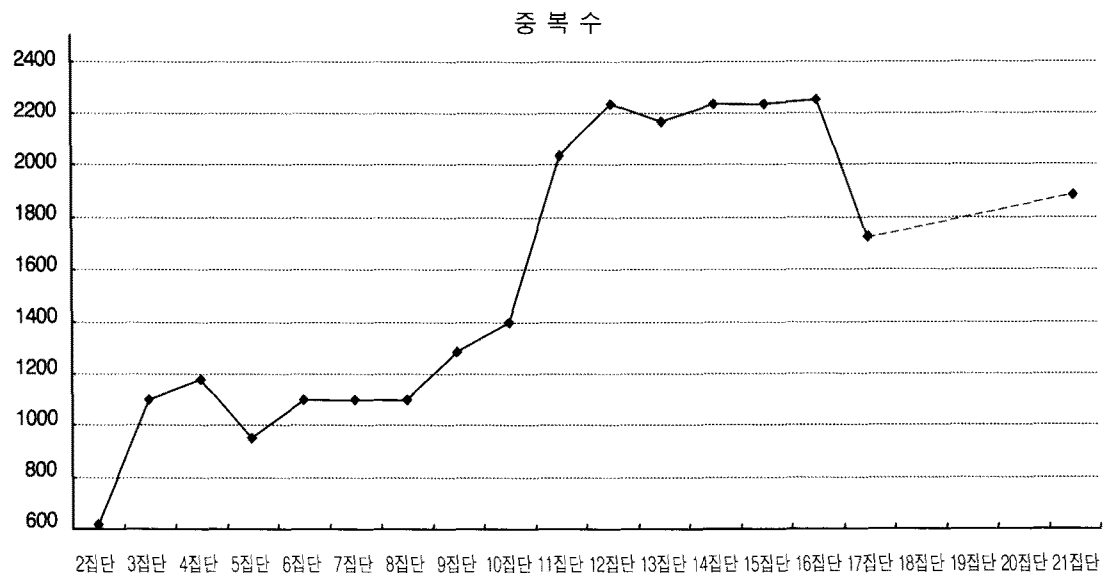
<그림 29>는 데이터베이스 구조에 따른 중복 수의 그래프이다.

10집단에 11집단으로 구조화 될 때, 데이터의 중복이 크게 나타남을 알 수 있다. 또한 16집단에서 17집단사이에서는 중복이 크게 줄어듦을 알 수 있다. 그림 26에서 10집단에서 11집단의 평균검색레코드 수의 감소 폭이 적은 것은 이러한 중복 현상이 영향을 미친 것으로 볼 수 있다. 또한 17집단에서 검색레코드 수의 감소 폭이 비교적 크게 나타남을 알 수 있다. 따라서 이 집단에도 소모시간은 어느 정도 감소하나 10집단의 소모시간보다는 높은 것으로 나타난다. 즉, 11집단에서 평균검색레코드 수는 적은 감소를 보이며, 열람을 고려해야 하는 데이터베이스가 하나 늘어남으로써 검색소모시간이 10집단에 비해 증가하게 되는 것이다. 결과적으로, 군집분석에서의 설명력을 기준으로 볼 때 13개의 카테고리분류가 이상적이었으나, 데이터의 중복, 시간소모

	1 집단	2 집단	3 집단	4 집단	5 집단	6 집단	7 집단	8 집단	9 집단	10집단	11집단	12집단	13집단	14집단	15집단	16집단	17집단	21집단
카테고리A	12052	4703	4418	4418	4098	3450	3450	3450	1693	1693	1693	1693	1693	1693	1693	1485	1485	1485
카테고리B		7963	5433	2441	2264	2264	2264	2247	2247	2247	1827	1827	1827	1827	1827	1827	846	257
카테고리C			3297	5433	4809	4809	4809	4809	4809	3602	1144	3602	3602	1144	3602	3602	229	229
카테고리D				938	938	938	493	493	493	493	3602	410	410	3602	410	410	225	225
카테고리E					892	812	812	812	812	812	410	812	812	410	812	812	1373	1373
카테고리F						876	876	876	876	876	812	876	796	812	796	796	3602	89
카테고리G							445	445	1947	1947	1315	445	445	1219	445	445	410	500
카테고리H								17	445	445	876	17	17	96	17	17	812	3491
카테고리I									17	17	1947	1487	1487	796	1487	1487	1219	410
카테고리J										1315	445	1315	1219	1487	1219	1219	96	613
카테고리K											17	1144	1144	656	1144	1144	796	320
카테고리L												656	656	445	601	601	1487	1219
카테고리M													113	17	80	80	601	96
카테고리N														80	96	96	445	148
카테고리O															55	55	55	796
카테고리P																225	17	1487
카테고리Q																	80	601
카테고리R																		445
카테고리S																		55
카테고리T																		17
카테고리U																		80
전체레코드수	12052	12671	13148	13230	13001	13149	13149	13149	13339	13447	14088	14284	14221	14284	14284	14301	13778	13936
중복수	0	619	1096	1178	949	1097	1097	1097	1287	1395	2036	2232	2169	2232	2232	2249	1726	1884

주) 비고: 중복수 = 집단(a)의 전체레코드 수 - 단일집단의 전체 레코드 수

<그림 28> 각 집단별 데이터의 분포 수 및 중복 수



<그림 29> 데이터베이스 구조에 다른 중복 수 그래프

의 결과 값을 기준으로 하였을 때는 10집단이 가장 이상적인 카테고리 분류로 나타났다.

Ⅵ. 결 론

본 연구는 기존의 연구와 실제 예제 데이터의 분석을 통해 MIRAS를 위한 다중플랫폼의 데이터베이스 재구조화를 구현하였다. 기존의 검색 시스템들은 각각 상이한 기준으로 카테고리를 분리하며, 관리하고 있다. 이렇듯, 명확하지 않은 분류 환경에서 MIRAS는 기존의 검색 시스템의 카테고리들을 통합하여 다중플랫폼의 데이터베이스로 재구조화 한다. 따라서, 효과적인 검색 시스템의 환경을 제공한다. 이러한 MIRAS의 데이터베이스 구조는 은닉시스템으로 존재한다. 또한, 실시간 데이터의 갱신과 학습을 통하여 최적의 분류 집단 수는 수정된다.

본 연구의 결과로는 각 집단별 및 요구 재현률에 따른 검색레코드 수와 열람데이터베이스 수의 관계를 규명하였다. 집단이 증가할수록 열람데이터베이스 수는 증가하지만, 본 실험 환경에서는 12집단에서 그 증가율이 떨어짐을 알 수 있었다. 또한, 검색레코드 수는 집단이 증가할수록 감소하나, 본 실험환경에서는 11집단에서부터 그 감소율이 떨어지며, 17집단에서는 다시 비교적 높은 감소율을 보인다. 이는 데이터 중복에 영향을 받은 것으로서, 각 집단별 데이터베이스의 레코드 분포를 분석한 결과 10집단에서 11집단으로 구조화될 때 레코드의 중복이 크게 나타남을 알 수 있었다. 또한, 17집단에서 중복이 다시 감소함을 알 수 있었다. 이러한 분석을 통해, MIRAS에서 재구조화된 데이터베이스는 검색소모시간을 단축시킴으로써 시스템의 객관적 성능을 향상시킨다.

또한, 체계적으로 재구조화된 다중플랫폼의 데이터베이스를 가지는 MIRAS는 사용자의 요구정보에 적절한 데이터베이스의 열람과 정보추출 방법을 제시한다. 즉, 사용자 질의에 불필요한 데이터베이스의 열람이나 레코드의 검색을 줄임으로서 효과적인 검색을 할 수 있는 기반을 구축할 수 있다. 본 시스템에서는 요구 재현률 80% 이상 및 70%이상시, 10집단에서 가장 효과적으로 데이터베이스를 재구조화 할수 있음을 알 수 있었다. 또한, 재현률 90% 이상을 요구시에는 단일플랫폼의 형태가 가장 효과적인 데이터베이스 구조라는 결과를 얻었다. 이는 사용자가 90%이상의 재현률을 요구하지 않는 검색 상황 즉, 대량의 선택 대안에서 몇 개의 대안을 선택하는 정보검색 과정에서 시스템 내의 검색소모시간을 단축시킬 수 있음을 나타낸다.

이러한, MIRAS는 각종 정보검색서비스간의 상호연계기술과 함께 보다 새로운 정보검색서비스를 구축할 수 있는 통합 데이터베이스의 재구조화 방안을 제시한다. 그리고, 정보화 시대의 각종 정보검색기술, 온라인 쇼핑몰의 제품분류 검색 등의 응용 기술로 중요한 부분을 차지한다고 할 수 있다.

본 연구는 구현하는 시스템의 환경에 따라 레코드 검색소모시간과 데이터베이스의 열람소모시간에 다소 차이가 있을 수 있다. 또한, 전반적인 시스템 소모시간은 관련도 및 사전 카테고리 구조학습방법, 적정 집단 수 결정의 방법적 개선으로 단축시킬 수 있다. 그러나, 본 연구는 MIRAS에서의 다중플랫폼의 데이터베이스의 재구조화에 관한 연구로서 이러한 요인은 고려하지 않았다.

또한, 이 연구는 명확히 MIRAS의 품질(Quality)을 평가한 것이 아니다. 다시 말해 검색소모시간의 비용과 정보검색의 재현률, 정보 품질 등의 가치 평가는 최종사용자에 의해 결정된다. 따라서, MIRAS의 품질 평가는 사용자의 주관적인 판단에 기인한 여러 요인을 포함하여 고려하여야 한다. 본 연구에서는 이러한 요인을 제외한 MIRAS의 객관적 성능과 관련되는 각 재현률 요구시, 집단별 검색소모시간, 데이터 중복의 자료만을 고찰하였다.

〈참 고 문 헌〉

[1] 신봉기, 김영환, "웹에이전트," *정보과학회지*, 제15권, 제3호, 1997.

[2] 최용석, "분산된 웹 데이터베이스에서의 정보검색 신경망 에이전트," 서울대학교 이학박사학위논문, 서울대학교대학원, 2000, pp. 1-144

[3] Blosseville M. and Hebrail G. and Monteil M. and Penot N., *Automatic Document Classification: Natural Language Processing, Statistical Analysis, and Expert System Techniques Used Together*, SIGIR'92, 1992, pp. 51-58.

[4] Cristina Bicchieri and Martha E. Pollack, and Carlo Rovelli, "The Potential for Cooperation among Web Agent," *Int. J. Human-Computer Studies*, 48, 1998, pp. 9-29.

[5] Gerald Hubl and Valerie Trifts, "Consumer Decision Making in Online Shopping Environments: The Effects of Interactive Decision Aids," *Marketing Science 2000 INFORMS*, Vol. 19, No. 1, 2000, pp. 4-21.

[6] Gravano, L. and Garcia-Molina H., "Generalizing GlOSS to Vector-Space Databases and Broker Hierarchies," In Proc. of the 21th Int. Conf. on Very Large Data Bases (VLDB), 1995, pp. 78-89.

[7] Howe, A. and Dreilinger, D., "Savvy Search A Metasearch Engine That Learns Which Search Engines to Query," *AI Magazine*, 18(2), 1997, pp. 19-25

[8] Hsiangchu lai and Tzyy-ching Yang, "A System Architecture for Intelligent-Guided Browsing on the Web," *HICSS*, 4, 1998, pp. 423-432.

[9] Kahle B. and Medlar A., "An Information System for Corporate Users: Wide Area Information Servers," *The Interoperability Report*, 5, Nov. 1991, pp. 2-9.

[10] Mauldin M. and Leavitt J., "Web Agent Related Research at the Center for Machine Translation," *in Proceedings of ACM SIGNIDR-94*, Aug 1994.

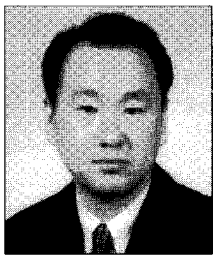
[11] Michael F. Schwarz, Alan Emtage, Brewster Kahle and Neuman B.C., "A Comparison of INTERNET Resource Discovery Approaches," *Computer Systems*, 5(4), 1992, pp. 461-493.

[12] Neuman B.C., "The Prospero File System: A Global File System Based on the Virtual System," *Computing Systems*, 5(4), 1992, pp. 407-432.

[13] Salton G., "The SMART Retrieval System Experiments in Automatic Document Processing," *Prentice-Hall*, Inc., Englewood Cliffs NJ, Chapters, 1971, pp. 14-17.

[14] Salton G. and McGill M.J., "Introduction to Modern Information Retrieval," *MacGraw-Hill Computer Science Series*, New York: McGraw-Hill, 1983.

◆ 저자소개 ◆



신창훈 (Shin, Changhoon)
현재 한국해양대학교 물류시스템공학과 부교수로 재직 중이다. 서울대학교 경영학과에서 학사 학위를 취득하였고, 한국과학기술원 경영과학과에서 석사 학위를 취득하였다. 한국전자통신연구소 연구원(1987~1990)에서 근무하였고, 한국과학기술원 산업경영학과에서 박사 학위를 취득하였다.

류병무 (Ryu, Pyungmu)
현재 (주) SLS 기획팀에 재직중이다. 한국해양대학교 물류시스템공학과에서 학사와 석사 학위를 취득하였다.

◈ 이 논문은 2002년 4월 29일 접수하여 1차 수정을 거쳐 2002년 11월 4일 게재확정되었습니다.