The Journal of Korean
Institute of CALS/EC
Vol 7, No. 2, June 2002

전문용어기반 eDocument 관리 방안에 관한 연구

김명옥*

A Study on eDocument Management Using Professional Terminologies

Myong Ok Kim

Abstract

Document retrieval (DR) has been a serious issue for long in the field of Office Information Management. Nowadays, our daily work is becoming heavily dependent on the usage of information collected from the internet, and the DR methods on the Web has become an important issue which is studied more than any other topic by many researchers.

The main purpose of this study is to develop a model to manage business documents by integrating three major methodologies used in the field of electronic library and information retrieval: Metadata, Thesaurus, and Index/Reversed Index. In addition, we have added a new concept of eDocument, which consists of metadata about unit documents and/or unit document themselves. eDocuemnt is introduced as a way to utilize existing document sources. The core concepts and structures of the model were introduced, and the architecture of the eDocument management system has been proposed. Test (simulation) result of the model and the direction for the future studies were also mentioned.

***Keyword** : document management, document retrieval, thesaurus, metadata, reversed index*

* 이화여대 경영대학

1. 서론

워드프로세서로 제작·출력된 문서를 처리하고 관리하던 사무자동화(office automa-tion)시대로부터 발전하여 초고속 정보 통신망과 네트워크 컴퓨터를 기반으로 하는 21세기 전자거래 기업 환경으로 진입하면서, 오늘의 사무문서관리는 상대방 컴퓨터로 전자문서를 직접 송수신하고 전자결재를 할 수 있는 단계에 이르렀다. 전자거래 환경의 확산과 지속적인 발전은 사무문서관리 분야에도 큰 변화를 가져올 것이다. 인트라넷과 인터넷을 통해 조직 내 모든 문서가 공유될 것이며, 통합 다큐먼트베이스 (document - base) 가 형성되어 궁극적으로는 종이 없는 사무환경이 조성될 것이다. 또한, 종이 없는 사무 환경에서 데이터베이스와 다큐먼트베이스 관리에 대한 전문 지식과 기술은 일반 사무관리직 종사자들이 당연히 갖추어야 할 필수 자격 요건으로 부각될 것이다.

그러나, 최종사용자 그룹 중 가장 큰 비중을 차지하는 일반 사무직 종사자들 가운데 상당수가 컴퓨터로 자동 처리해야 할 성질의 업무를 아직도 수작업으로 처리하고 있는 실정이다. 최근 한국비서협회가 실시한 설문조사 결과에 의하면, 일반 사무직 근로자들의 업무시간 구성 중 가장 많은 시간을 소요하는 업무를 순서대로 나열하였을 때, 문서작성업무 15%, 계획/대안 분석 및 평가업무 14%, 자료수집업무 13%, 문서검색 8%, 문서 정리, 분류, 보관업무 9% 등으로 문서의 작성, 검색, 저장과 관련된 활동에 총 32%에 달하는 시간을 소요하고 있는 것으로 나타났다. 또한 사무직 근로자들은 업무수행 활동 중 효과적 수행이 가장 어려운 활동으로 문서관리 업무를 꼽았다. 한편, 현직 비서직 종사자 200명을 상대로 조사한 결과에서도 문서관리에 가장 많은 시간과 노력을 투자하는 것으로 나타났다. 특히 문서의 효율적 검색에 관해서는 전산시스템의 효과가 거의 미치지 못하고 있으며, 전통적 파일링, 문서 캐비닛 사용, 그리고 디스크 파일 관리 등이 아직까지 문서관리의 보편적 수단인 것으로 드러났다. 이렇듯 왕성한 전자거래 시대의 도래를 목전에 두고도 자금까지 일반 사무 환경에서는 문서와 양식의 정리, 분류, 찾기와 같은 사무문서관리 작업은 거의 100% 재래적인 비전산화 방식으로 이루어 지고 있는 상황이다 (한국비서협회, 2000).

기업 내에서의 문서의 흐름은 바로 정보의 흐름을 뜻하기 때문에, 위와 같은 문제는 한 조직의 효율적 운영 및 관리에 큰 걸림돌이 될 것이므로, 문서관리 자동화를 통한 일반 사무관리직에서의 자원 절약과 질 높은 서비스/결과 산출, 그에 따른 직무만족의 증대는 매우 큰 의미를 지니고 있다.

본 연구의 목적은 특정 전문직의 전문용어를 기반으로 한 효율적 사무문서관리 방안의 모색에 있다. 메타데이터와 전문용어집을 이용하여 가상통합문서를 생성하고 관리할 수 있는 반자동 시스템의 설계와 알고리즘 시뮬레이션이 본 연구의 핵심이다. 시뮬레이션은 국제회의 관련 직무를 분석하여 선정된 500단어 전문용어집과 5인의 전문가 패널에 의해 이루어졌으며, 사무문서관리에 있어 경제성과 정확성 측면에서 긍정적인 결과를 도출했다.

2. 사무 문서 관리

사무 문서라 함은 보통 사무실에서의 고유 업무 이외에 일반적인 업무에서 발생하는 문서, 예를 들면, 대외문서, 보고서, 결재서류, 각종 대장 등의 행정 업무상 발생하는 일반 서류들을 의미한다 (심재균, 1998). 이들은 전 산업분야에 걸쳐 일반사무 업무라면 거의 필연적으로 발생되는 문서이며, 대규모 조직인 경우에는 문서관리 규정을 두어 발생에서부터 유통 (주로 결재), 배포, 보관, 이관, 폐기 등에 이르는 문서의 생명주기(Life Cycle) 및 보관관리 체계를 정의하고 있다. 그러나 조직이 작거나, 경영층의 특별한 의지가 없으면 전산화가 잘 이루어 지지 않는 부분이며, 그 효과에 있어서도 잘 인식되지 않는 영역이다. 그러나 어떤 조직에서건 조직활동 중 발생하는 문서들은 중요한 자원으로 다루어져야 하며, 특히 결재 과정을 거치는 문서들은 여러 사람의 의지가 들어 있는 정보의 보고이며 필수 관리 대상이 된다.

Thanos(2000)는 사무문서를 종이문서, 이미지문서, 수정 가능한 문서, 그리고 복합문서로 나누었다. 종이문서(Paper Document)란 문서 생성 과정을 통해 최종적으로 출력물 형태로 전환되어 종이의 형태로 존재하고 있는 문서로 스캐너, 팩스 또는 OCR 등의 기술을 이용해 디지털 정보로 전환시킬 수 있다. 이미지문서(Image Document)는 종이문서를 스캐너 등을 이용하여 실물보관 및 조회 등의 목적으로 이미지화 시킨 문서이다. 수정 가능한 문서(Editable Document)란 워드프로세서, 스프레드시트 등의 문서생성도구를 통해 작성되어지는 문서로 좁은 의미의 전자문서이다. 수정 가능한 문서의 특징은 문서작성 도구를 통해 신규 창출 또는 수정 등이 용이하다는 점이다. 복합문서(Compound Document)는 HTML 등의 문서를 의미하는 것으로 네트워크를 통해 정보를 직접 공급 받아 재사용 가능하며, 이미지 문서, 수정 가능한 문서 등을 모두 포함하여 텍스트 뿐만 아니라 이미지, 차트 등의 다양한 정보를 표현할 수 있는 문서를 의미한다.

문서정보는 대개 일정한 문서 생명주기에 의해 철저한 색인 및 파일링 작업을 거친 후 보관, 관리되어야 하나, 업무의 폭주와 대부분 눈 앞에 놓인 일상 업무를 먼저 처리하게 되는 습관으로 인해 제대로 행해지고 있지 못한 실정이며 문서관리가 어려운 직접적인 요인이 바로 이것이다(Shelley, 1995). 체계적인 문서관리시스템을 통해 조직 내 중요한 정보자원을 효과적으로 관리하기 위한 적절한 시스템 및 관리통제, 보안체제 구축을 이루고 무엇보다도 문서관리의 조직 내 표준화를 이루는 것이 시급한 일이다.

웹 기반 문서관리에서의 가장 중요한 과제도 일반 문서관리에서와 같이 빠르고 정확한 저장과 검색이지만, 정보의 바다라고 불리는 인터넷의 등장으로 이제는 지식이나 문서 그 자체 보다는, 빠르고 정확하게 원하는 내용을 습득하고 찾아내는 문서검색능력이 더욱 중요하게 되었다. 그 이유는 매3 ~ 4개월마다 인터넷을 통하여 얻을 수 있는 정보가 두 배로 늘고 있기 때문이다. 따라서 각 조직은 대내적으로는 문서의 보관공간 절약에 따른 사무환경 개선은 물

론 정보의 빠르고 정확한 검색과 대외적으로 업무 지원이나 각종 서비스의 질적인 향상을 위해 새로운 문서관리 체계가 필요하게 되었다. 다양한 디지털 정보시대에 그에 상응하는 전자 문서관리 체계가 확립된다면 문서검색, 문서공유, 문서관리의 일관성, 서류의 중복저장 제거, 문서전송 용이, 문서 버전 관리 용이, 문서표준화 등과 같은 효율성을 기대할 수 있다 (정영미, 1993).

Sacks-Davis(1992)는 문서 구조화 모델과 문서 검색 모델을 제시하고 통합된 문서관리 시스템을 클라이언트 서버 방식으로 구현하였다. 기존의 문서관리 시스템에 OMT (Object Modeling Technique) 방법을 적용하여 객체지향적 문서관리시스템을 설계하고 구현하였으며, 시스템의 확장성과 유지보수, 환경변화에 쉽게 대응할 수 있는 문서관리시스템의 중요성을 강조하였다. 문주영(1998)은 비서직 종사자들의 효율적인 사무문서 관리를 위한 방안으로 문헌정보학 분야에서 문헌을 기술하거나 문헌의 목록을 작성하는 데 쓰이는 메타데이터를 기반으로 하는 문서모델을 제안하고 시스템의 프로토타입을 구현하였다. 장미경(1998)은 비서직 종사자들의 효율적인 사무관리를 위해 시소러스와 색인데이터베이스를 기반으로 한 웹 기반 문서관리시스템을 설계 구현하였다. 최수진(2001)은 지식공유를 위한 가장 효율적인 수단으로의 자동문서관리에 대한 기초 연구 결과를 소개했다.

서지 데이터 관리, 사무정보관리, 비서학 등의 분야에서 관심을 갖고 있으며, 선행 연구들은 물리적으로 입력 저장되어 실존하는 문서 개체들 만을 그 관리 대상으로 삼고있는 것이 공통적 특성이다.

3. 문서 검색

내용기반 문서 검색에 큰 관심을 갖고 많은 연구를 진행하고 있는 분야가 전자도서관이다. 전자도서관은 멀티미디어 정보의 디지털 처리, 저장 및 통신을 위한 하드웨어와 디지털 정보의 수집, 저장, 분류, 검색, 그리고 분배를 위한 소프트웨어가 결합된 정보기반 구조라고 할 수 있다 (Subrahmanian, 2000).

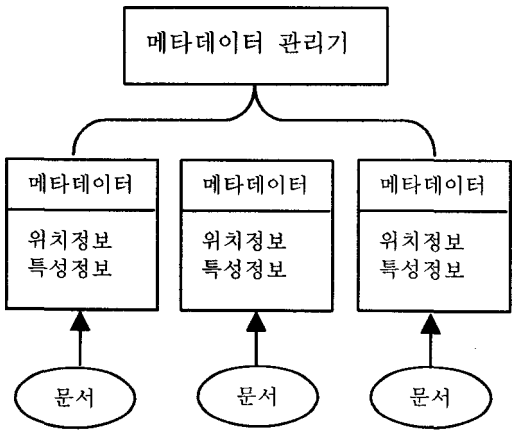
전자도서관은 대량의 정보 모델 구축과 검색을 기본으로 하기 때문에 성공적인 구현을 위해서는 무엇보다도 데이터베이스시스템 기술과 정보 검색 기술을 반드시 필요로 한다. 지리적으로 분산되어 있는 데이터베이스를 사용자가 하나의 통합된 데이터베이스처럼 사용하도록 하는 연합데이터베이스 (Federated Database System) 방식이어야 하며, 검색 역시 통합데이타베이스 내에서 분산된 정보에의 자유로운 접근을 허용하는 통합정보 검색이어야 한다(Yates, 1996). 전자문서 관리를 위한 문서모델 개발에 필요한 기초 개념으로 전자도서관의 구현의 핵심 기술과 도구를 살펴본다.

3.1 메타데이터

메타데이터 (Meta Data: Data about data)란 데이터에 관한 특성정보를 뜻한다. 메타데이터는 수록된 데이터의 내용, 품질, 조건 및 그 데이터가 갖고 있는 특징을 알려 주는 데이터로 정의한다(Shelley, 1995).

또한 Gurda, Danielsen과 Hemstead (1995)은 메타데이터를 데이터베이스에 접근하기 위해 미리 그 데이터의 성질, 내용, 품질, 조건, 특성을 기술하는 데이터로 정의하고 있으며 데이터의 의미를 기술하거나 대표하는 기능을 한다고 하였다. 전자도서관의 검색시스템이 메타데이터의 대표적인 사용 예가 된다. 현재 전자도서관에서 매우 활발히 논의되고 있는 것이 다양한 구조와 형식을 갖는 정보에 대한 특성정보 및 분산 저장된 디지털 정보의 식별 방법에 관한 표준 모델의 마련이다. 사용자에게 찾고자 하는 자료의 위치를 정확히 추적하는데 필요한 정보를 제공해 주며, 아울러 그 자료가 정말 양질의 데이터인지를 파악하는 것을 가능하게 해준다 (Schatz, 1997). 이러한 메타데이터는 시간과 비용의 낭비를 줄이고 불필요한 송수신 과정을 간소화 시킴으로써 정보 유통의 효율성을 제고시킬 것이며, 더욱이 모든 사무 정보가 일정한 표준에 의해 구축될 경우 그 이용으로 인한 효과는 매우 클 것이다.

여러 종류의 메타데이터 모델이 존재하나, 공통된 속성으로 제목, 주제, 요약, 관계, 보안, 날짜, 작성자, 출판사, 포맷과 같은 것들이 있다(Shelley, 1995). 1995년 관련 전문가들이 모여 인터넷과 같은 네트워크 환경에서 전자문서를 처리하는데 필요한 최소한의 메타데이터를 15개 요소로 정하고 이를 더블린 코어라 명명하였다. 메타데이터가 정보의 관리에 사용될 경우 이용자들은 다음과 같은 효과를 얻을 수 있다 (Gurda, 1995): (1)메타데이터는 작성한 실무자가 바뀌더라도 변함없는 데이터의 기본 체계를 유지하게 함으로써 일정한 시간이 지나도 일관성 있는 데이터를 이용자들에게 제공할 수 있고, (2)데이터를 목록화 하기 때문에 사용하기에 편리한 정보를 제공하며, (3) 정보 공유의 극대화를 도모하여 데이터의 원활한 교환을 지원하기 위한 틀을 제공한다.



<그림1> 메타데이터 개요

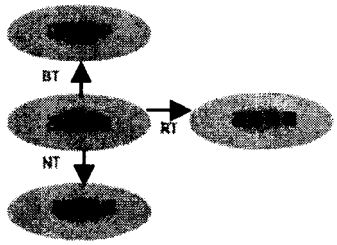
물리적 문서 한 개당 메타데이터 한 세트를 작성하여 저장하고, 그렇게 모여진 메타데이터 세트를 메타데이터 관리기가 통제한다 (그림1). 단순 필터링이나 인덱싱 기법에 의해 입력된 여러 가지 종류의 키워드와 연관된 문서 정보를 입수하고 저장 위치를 확인하거나, 문서의 특성정보 중에서 관계(Relation) 정보를 이용하여 현재의 문서와 논리적으로 연관되는 다른 문서들로 접근한다.

3.2 시소러스

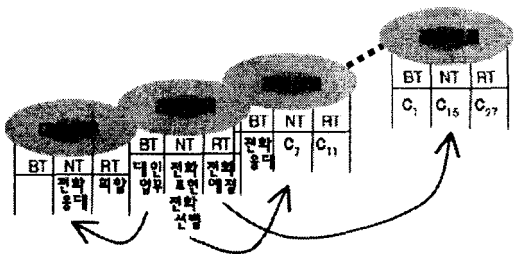
시소러스(Thesaurus)는 기능적인 관점에서 자연 언어를 보다 통제된 시스템 언어로 변환하기 위한 어휘 통제장치라고 정의할 수 있으며, 구조적인 관점에서는 상위 및 하위 개념사이의 전후 관계를 명백하게 하기 위하여 의미상 관련 있는 용어들을 모아놓은 동적 어휘집이라고 정의한다. 일반 사전은 용어의 의미를 모를 때 사용하는 도구인 것에 반해, 시소러스는 의미(개념)는 이해하나 그에 해당하는 용어를 모를 때 사용하는 수단이다(Toshio, 1995). 찾고자 하는 정보가 무엇인지 알고 있어도 그 정보에 대한 정확한 검색 용어를 알지 못하면 정보의 입수가 불가능하다. 정보검색을 위한 최초의 시소러스는 1959년 DuPont사의 시소러스가 그 효시를 이루며, 현재는 전통적인 도서관의 주제명 목록에서 시작하여 컴퓨터 정보검색을 위한 정보생산자, 입력자, 사용자 사이의 용어통일의 역할을 수행하는 용어집으로 개념이 확장되어 가고 있다. 여러 응용 분야에서 시소러스의 기본 기능은 동일하나 시소러스의 이용 방법과 그에 거는 기대 효과는 다른 양상을 보이고 있다.

자연어에는 다른 어휘보다 해당 주제에 가까운 것, 보다 넓은 개념, 보다 애매한 개념들이 섞여있고 서로 밀접하게 관련되어 있다. 다양한 유사어나 동의어, 동음이의어의 다의성 문제, 문맥의 모호함, 개념사이에 존재하는 계층관계 등과 같은 혼란 속에서 일반사용자가 아무 도움없이 이런 개념을 찾아내고 식별한다는 것은 거의 불가능하다고 본다 (Davis, 1996).

시소러스는 여러 개념들을 그들 사이의 관계로 표현한다 (그림2). 하나의 개념은 통제어 한 개와 그에 종속된 유사어의 집합으로 정의되고, 각각의 통제어는 상위관계(BT: Broader Term), 하위관계(NT: Narrower Term), 그리고 연관관계(Relational Term)를 표현하는 통제어들과 복합적 계층구조를 맺고 있다(Crouth, 1990). <그림 3>과 같은 시소러스 구조는 사용자가 사용한 자연언어를 색인 작성과 검색에 사용할 통제어휘로 번역하는 수단을 제공하고, 색인어 부여시 일관성을 보증하며, 용어간 의미관계를 지시하고, 그리고 문서 탐색을 위한 보조 도구가 된다.



<그림2> 개념 관계



<그림3> 시소러스 구조

정보검색 분야의 연구에 있어서 특징적인 사항중의 하나는 많은 경우에 직관적 통찰에 의해 개발된 검색 기법들이 검색 효과의 향상을 가져오지 않았고, 그 한 예가 시소러스를 사용함으로써 기대했던 만큼의 검색 효과의 향상을 얻는데 실패했다 (이준호, 1998). 이러한 문제를 수정보완 하기 위한 한 방법으로 본 연구에서는 내용기반 시소러스, 즉, 특정 직무 관련 전문용어집을 검색의 핵심 도구로 삼는다.

3.3 색인

정보검색 과정은 축적과 검색으로 구분되며 축적을 위해서는 색인작업이 필요하다. 색인이란 정보사용자가 쉽게 접근할 수 있도록 정보원에 포함된 정보내용을 쉽게 탐지할 수 있는 소재지시기호를 달아 일정한 순서로 배열하는 것을 말한다. 색인어를 기반으로 하는 정보검색모델에서 질의어안에 포함된 색인어들은 서로 독립적이어서 사용자가 도메인에 대한 전문지식이 없거나 정확한 색인어를 모를 경우 자신이 필요로 하는 정보를 효율적으로 기술하기 어렵다 (Berchtold, 2000). 따라서 사용자는 자연어를 이용하여 정보요구를 표현하고 검색시스템은 이를 분석하여 사용자에게 정보를 제공하므로 검색이 적절히 이루어지지 않을 수 있다. 이때 시소러스는 색인과정에서 개념이 다른 용어로 표현되는 유사어를 통제하는데 이용된다 (Davis, 1996).

색인은 관점에 따라 여러 가지 형태로 분류할 수 있는데 방법에 따라 수동색인과 자동색인, 이들을 혼용해서 쓰는 반자동색인으로 구분할 수 있다. 수동색인에서 주제분석은 일반적으로 색인자에 의해서 수행되고 시소러스를 이용해서 색인어를 부여하는 방식이 쓰인다. 반면, 자동색인은 대량의 문헌정보 데이터베이스로부터 필요한 정보를 빠르게 추출하는 데는 필수적인 기술이지만, 현재까지의 자동색인기술은 필요한 주제어를 추출하는 동시에 불필요한 주제어도 추출하게 되지만 문장에서 나타내고자 하는 개념을 추출하지 못하게 되는 문제점이 있다. 이상적인 자동색인은 대상문헌에서 주제어를 추출하고 필요하다면 본문에서 사용되지 않은 단어를 추출해야 하며, 그 단어를 통제어 형태로 변환시킬 수 있어야 한다. 자동색인에서 시소러스의 역할은 유사어의 경우 이를 통제어로 바꾸어 색인어를 일치시키고 여러 가지 색인기법을 사용하여 주제어를 선정할 때, 그 단어가 시소러스 용어인지 아닌지를 확인하고 시소러스 통제어일 때 가중치를 많이 부여해 검색효율을 높이는 것이다.

사용자가 원하는 정보의 특성을 잘 알거나 정확한 키워드를 입력하는 경우는 드물며, 자신이 원하는 정보에 관한 의미적, 개념적 정보만을 가지고 있는 경우가 대부분이다. 특히 멀티미디어 정보를 다루어야 하는 전자도서관의 경우 내용 기반 검색이 매우 중요하다 (Schatz, 1997). 멀티미디어 정보 중에서 전문(full text)에 관한 내용 기반 검색은 정보 검색 시스템 분야에서 계속 연구되어왔으며, 그림, 음성, 동화상 등 다른 형식의 멀티미디어 정보의 내용 기반 검색에 기초가 된다. 정보 검색을 위해서는 저장된 전문정보에 대하여 내용 기반 색인

이 마련되어야 하며, 이 색인은 단어의 출현 횟수를 기반으로 하는 자동 색인기법에 의하여 구축된다. 전문 이외의 다른 형태의 멀티미디어 정보에 대한 내용 기반 검색은 현재 활발한 연구가 진행 중이다 (Khoshafian, 1997: 김형주, 1999).

검색시스템에서 원하는 정보에 대한 색인어를 입력해 정보를 검색하는 것이 색인어 검색 방법이다. 저장시스템은 크게 지식문서를 저장하고 있는 문서DB와 문서에서 색인어를 추출한 색인DB로 나눌 수 있다. 색인DB는 시소러스를 이용하여 생성된, 문서의 색인어를 담은 파일이다. 이 데이터베이스에 대해 사용자는 시소러스를 통해 검색어를 구체화시키고, 시스템은 이에 적합한 정보를 검색하여 이용자에게 제공한다. 색인어 저장 시스템으로는 역색인 방식이 있는데 순차 저장된 원문들로부터 추출된 색인어를 키로 사용하여 해당 색인어들을 갖고 있는 문서정보를 빠르게 검색하는 것이다. 역색인 방식은 데이터에 대한 색인정보를 미리 구축해 놓고 원하는 색인어에 대한 질의가 입력되었을 때 미리 구축된 색인 정보를 이용해 신속히 정보를 찾아줄 수 있는 시스템으로 정보의 양이 증가하더라도 검색 속도가 크게 영향을 받지 않고 논리검색, 시소러스 검색 등 다양한 검색 기능 구현이 가능하며 현재 가장 널리 사용되고 있는 방법이기도 하다.

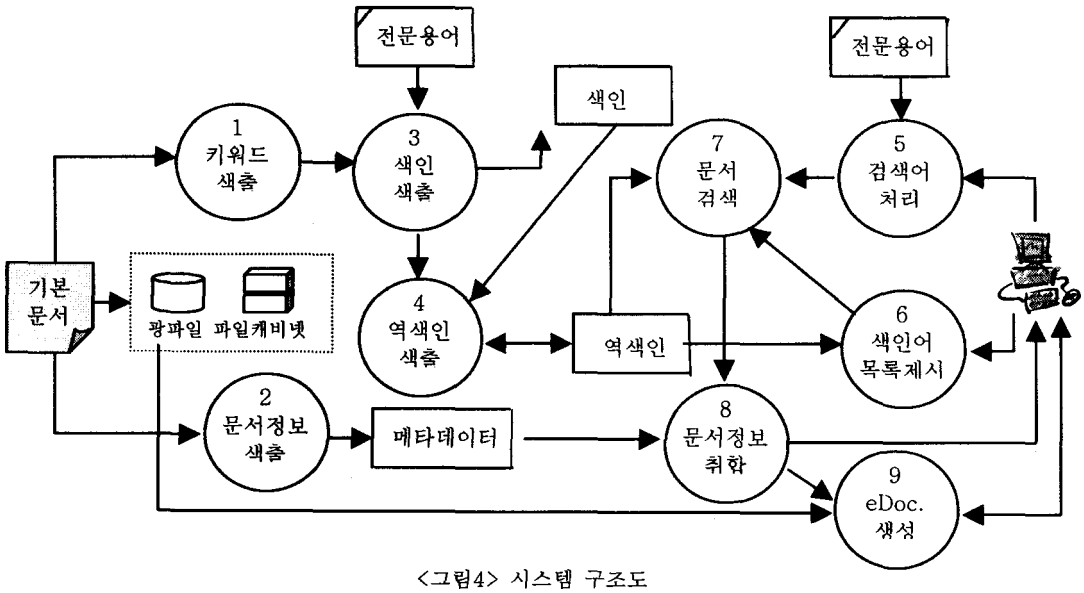
자동색인에 대한 긍정적인 연구결과 들이 있으나, 경제성과 멀티미디어 자료에 대한 유연성 관점에서는 아직 크게 향상된 효과를 기대하기는 어렵다.

4. 전문용어기반 eDocument 관리

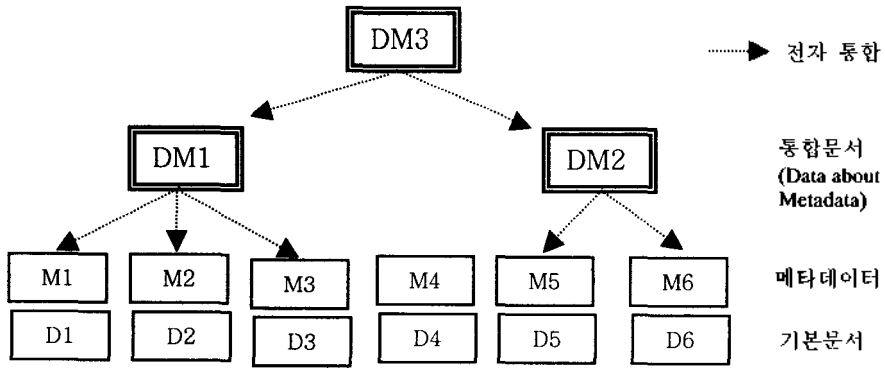
특정 전문직의 전문용어를 기반으로 한 사무문서관리 방안을 그림4와 같이 제시한다. 본 연구에서 사무문서라 함은 물리적으로 저장된 하위 수준의 기본문서와 기본문서들을 통합하여 형성된 상위 수준의 전자통합문서를 포함하는 광의의 문서를 (eDocument라 칭함) 의미한다 (그림5)

사무문서관리는 문서관리자가 여러가지 형태와 종류의 문서들을 그 내용을 기준으로 분리하여 저장하는 것으로부터 출발한다. 처리의 대상이 되는 문서를 기본문서 (unit document)라 칭한다. 기본문서로는 전통적인 텍스트 유형의 문서뿐만 아니라, 소리나 동영상 자료에 대한 메타데이터 관리가 이루어질 수 있기 때문에 멀티미디어 구성도 가능하다.

본 연구의 시스템(그림4) 에서는 기본문서의 내용을 점검하고 키워드를 색출하고, 제반 문서정보를 축출하여 요약하는 일은 문서관리자의 수작업으로 이루어진다. 문서의 내용을 이해하고 적합한 주요어 색출 작업을 자동화 하는 데에는 시스템의 안정성, 결과의 정확성, 그리고 무엇보다도 경제성의 측면에서 아직은 고비용 저효율의 패러다임을 벗어나지 못하고 있기 때문이다 (최기선, 2002: Ricardo, 1999).



<그림4> 시스템 구조도



<그림5> eDocument 개념

4.1 자료 구조

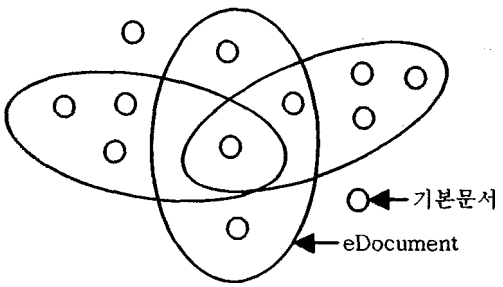
4.1.1 eDocument 구조

<그림 5>에서 기본문서 개체 D1, D2, D3, D4, D5, 그리고 D6는 각각의 해당 메타데이터 M1, M2, M3, M4, M5, 그리고 M6에 의해 요약된다. 개체 DM1은 기본문서 D1, D2, D3을 병합하여 생성된 가상통합문서이고, 개체DM2는 기본문서 D5와 D6를 통합한 가상통합문서이며, 개체 DM3는 DM2와 DM3를 다시 병합한 상위 가상통합문서이다 (그림6). 한 번 저장된 문서

개체의 재활용을 용이하게 하는 문서구조라는 데 큰 의의가 있다.

<그림 5>의 메타데이터는 기존의 Dublin Core 모델이 포함하는 15개의 항목 (Title, Subject, Description, Source, Language, Relation, Coverage, Creator, Publisher, Contributor, Rights, Date, Type, Format, Identifier)으로 구성되었고, 기본문서 개체를 정의하는 중심 내용이 된다.

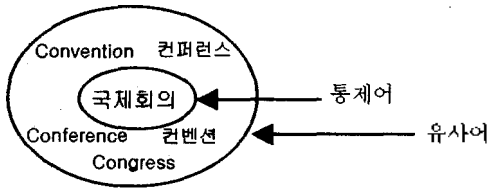
본 장에서 제안하는 문서 모델의 특성은 DM(Data about Metadata) 노드에 있다. 포맷의 구속에서 벗어날 수 있는 디지털 문서의 장점을 최대로 살렸으며, 기본문서의 메타데이터만을 병합함으로써 대규모 상위 가상문서의 생성을 꾀할 수 있다. 저장 매체에 저장된 기본문서에 대해서만 메타데이터를 생성하고 관리하는 것이 아니라, DM 또한 관리 대상으로 포함한다. DM의 구성 요소로 위치정보, 작성자 혹은 책임자, 작성 날짜, 키워드를 포함한다.



<그림 6> 기본문서와 eDocument

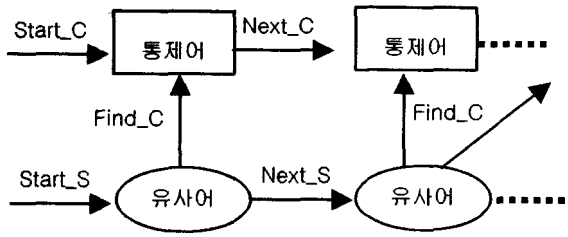
4.1.2 전문용어 DB 구조

앞 장에서 살펴 본 전형적인 시소러스의 구조를 아래 그림8과 같이 수정하였다. 통제어(그림7)는 특정 개념을 대표하는 용어로 다수의 유사어를 거느린다. 그러나, 문서 검색을 위해 사용자가 제시하는 주제어를 유사어로 간주하기 때문에, 주어진 통제어에 종속된 유사어를 색출하는 포인터 *Find_S* 는 고려하지 않기로 한다.



<그림 7> 통제어 예

유사어는 통제어와 비슷한 의미를 내포하는 용어로서 다수의 통제어에 귀속될 수 있게 했다. 사용자가 제시한 검색어를 유사어로 전환하고 *Find_C* 포인터를 따라 연결된 모든 통제어를 축출하는 것이 가능하다. 예를 들어, 사용자가 선택한 키워드가 '컨벤션'이라면 검색 결과로 통제어 '국제회의'를 축출하게 된다.



<그림 8> 전문용어 통제어/유사어 구조

4.2 저장 및 검색

본 시스템(그림 4)의 핵심 모듈들의 기능을 요약하면 아래와 같다.

4.2.1 키워드 및 문서정보 색출

키워드 자동생성에 따른 고비용 저효율의 문제를 극복하기 위한 대안으로 본 시스템에서는 기본 문서의 내용을 숙지하고 다수의 키워드를 색출하는 작업을 문서 관리자의 분석과 판단에 맡긴다. 아울러 문서의 위치정보, 책임자(부서), 그리고 작성날짜도 문서에 대한 메타데이터로 기록해둔다.

4.2.2 색인 색출

```
{input: document id D, a set of keywords
        K, thesaurus T;
 output: a set of control words C}

Let K to be a set of keywords selected
from the source document;
C = { };
For each k ∈ K,
    Follow the Next_C to find k in
        Start_C↑;
    If found, then C = C + k;
    Else follow Next_S to find k in
        Start_S↑;

        If k is found as s in Start_S↑,
        then C = C + Find_C(s)
        Else add k to Start_S↑ and
            Start_C↑
                C = C + k
Store D and C in 색인DB;
Return C;
Endfor;
Endprocedure
```

4.2.3 역색인 색출

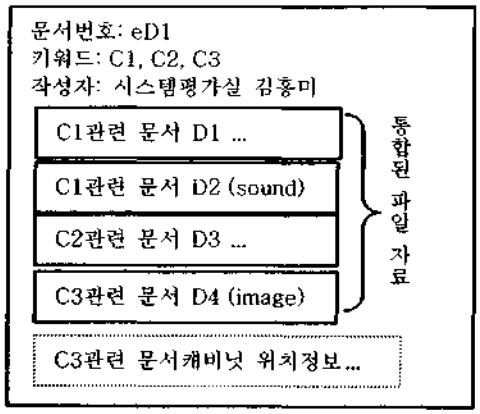
```
{ input: document id D, a set of control
         words C, 색인DB, 역색인DB;
 output: updated 역색인DB }
For each c ∈ C,
    Find all documents in 색인DB;
    Add c with documents into 역색인
DB;
Endfor;
```

4.2.4 검색어처리 및 색인어목록 제시

사용자가 키워드에 근거한 문서검색을 실시할 경우, 위에서 설명한 관리자 모듈의 색인 색출 과정과 흡사한 방법으로 통제어가 산출된다. 사용자가 키워드를 쉽게 선택할 수 있도록 역색인DB의 통제어 목록을 제시해 주는 것도 한 방법이 될 수 있다.

4.2.5 문서검색 및 문서정보취합

사용자가 제시한 키워드는 전문 용어집의 통제어 C로 전환되고, 역색인 DB로부터 통제어 C를 포함하는 모든 문서번호를 색출한다. 검색된 문서번호를 이용하여 저장된 문서의 위치정보와 특성을 파악하여 사용자에게 전달한다. 한편, 통제어 C와 관련 문서정보를 이용하여 eDocument가 생성된다 (그림9). 디지털 파일 형식으로 저장된



<그림 9> eDocument 예

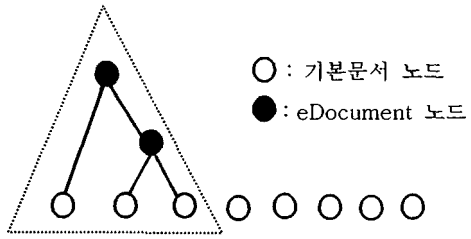
기본 문서들은 즉시 통합이 될 수 있는 데 반해, 계래 문서들은 그 메타데이터가 추가되어 전자통합문서로 정의되어 사용자에게 전달된다. eDocument 생성을 위해서는 사용자가 색인어 목록을 참조하여 키워드를 다수 선정하고, 선택된 통제어와 관련된 문서들을 취합할 수도 있다.

5. 시뮬레이션

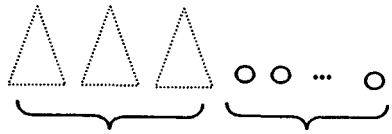
시스템 알고리즘 시험운영을 위해 Windows XP Professional 운영 체제와 MS Access XP 프로그램으로 환경을 구성하였다. 실험을 위해 국제회의 관련 어휘 500단어로 모형 전문용어집을 구성하였다 (부록).

기본문서로부터 키워드를 색출하기 위하여 문헌정보관리자 2인, 전문비서 2인, 국제회의기획 전문가 1인을 포함하여 전문가패널(panel)을 구성하였다.

국제회의기획 및 운영을 위한 사무문서들을 관리대상을 삼았고, 문서의 수를 10개, 30개, 50개, 그리고 100개의 네가지 사례로 나누어 시험을 하였다. 한편, 문서의 규모도 대(3000 단어 이상 포함), 중(1500 이상 3000 이하의 단어 포함), 소(1500 단어 미만 포함)의 3가지 종류로 구분하였다. 위의 4가지 사례 모두에 대중소 문서 크기 비율을 각각 4:3:3으로 동일하게 적용하였다 (그림10). eDocument 구조는 기본문서의 양적 증가와 관계없이 모든 사례에서 그 특성이 동일하게 유지되었다 (그림11).



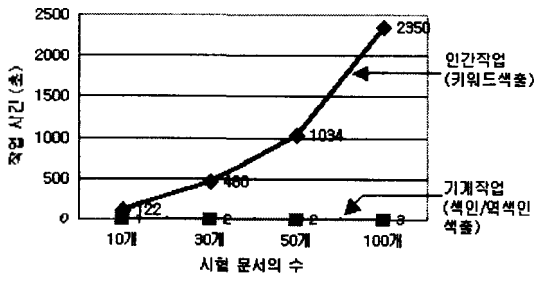
<그림10> eDocument 구조 예 (시험 문서10개)



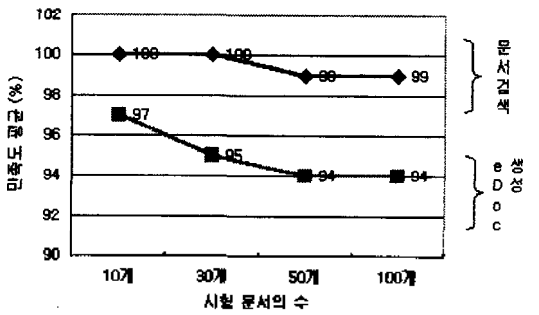
<그림11>eDocument 구조 예 (시험 문서30개)

각 시험 사례에 대하여 5인 전문가들이 기본문서로부터 키워드와 문서특성정보를 색출하여 입력하는데 걸린 평균 시간과 전문용어집을 참고하여 색인/역색인을 색출하는 기계 반응시간을 측정 비교하였다 (그림12). 문서의 수 가 증가함에 비례하여 인간 작업은 큰 폭으로 급등한 반면, 시스템은 실행에 있어 안정성을 보였다.

각 시험 사례에 대하여 문서검색과 eDocument 생성 결과의 정확성에 대한 전문가 패널의 만족도 평균은 그림13과 같다. 문서의 수가 증가하며 전체 만족도는 감소하는 경향이 있으나, 시스템의 안정적 실행 성능에 비해 인간작업 부문에서 소요되는 작업시간의 증가에 기인한 만족도 감소라고 판단된다.



<그림12> 작업시간 (인간 vs. 기계)



<그림13> 시스템 만족도

6. 결론 및 제언

특정 전문직 내에서의 사무문서를 통한 지식공유를 목적으로, 전문용어를 기반으로 한 효율적 사무문서관리 방안을 제시하였다. 문서의 핵심 내용을 자동 축출 하기위한 방법 모색에 많은 연구가 이루어지고 있고, 주목할 만한 결과도 눈에 띄고 있으나, 본 연구에서는 시스템의 편리성, 경제성, 안정성, 그리고 무엇보다도 정확성/만족도에 초점을 맞추었다.

단락(paragraph), 절(section), 장(chapter), 그리고 보고서(report)와 같은 텍스트 유형의 기본문서와 그들을 통합하여 구성된 eDocument를 그 관리대상으로 정했고, 모의실험에는 포함하지 않았으나, 기계작업으로는 그 내용을 자동 색출하기에 무리가 따르는 유형의 자료인 소리나 이미지와 같은 멀티미디어 자료도 관리가 가능하다. 멀티미디어 자료를 분석함에 있어 아직은 효율적인 인간작업을 적절히 포함시켰다.

국제회의관련 작무를 분석하여 용어 500개를 선정하여 전문용어집을 구성하였고, 이를 기반으로 본 연구 시스템에 대한 모의실험을 실시하였다. 문서의 수가 증가함에 따른 인간작업 시간의 증가추세로 인해 야기되는 문제에 비해, 문서검색과 eDocument 생성에 있어 그 정확성에 대한 안정적인 실행과 사용자 만족도는 주목할 만한 결과라고 판단된다.

본 연구 결과의 일반화를 위해서는 참가한 전문가의 수 증가, 기본 문서의 수 증가, 문서 내용의 다양화, 전문용어집의 내용 확대, 다양한 전문 영역에의 적용 등과 같은 확대 시도를 필요로 한다.

참 고 문 헌

[김형주, 1999] 대규모 이미지 데이터베이스에서 고차원 색인 구조를 이용한 효율적인 내용 기반 검색 시스템, 정보과학회논문지(B), 제26권 제1호, 1999.

[문주영, 1998] 문주영, 메타데이터를 이용한 웹기반 문서관리 시스템 개발에 관한 연구, 비서학논총 제16권 제2호, 1998.

[심재균, 1998] 심재균, "EDMS 구축 사례 및 기대효과," EDMS 기술백서, 한국경제신문사, 1998.

[이준호, 1998] 이준호, 정보검색시스템 평가 및 테스트 컬랙션 개발, 정보과학회지, 제16권 제8호, 1998.

[장미경, 1998] 장미경, 비서직을 위한 시소러스와 색인데이터베이스를 활용한 웹기반 문서관리시스템의 설계 및 구현, 비서학논총 제16권 제2호, 1998.

[정영미, 1993] 『정보검색론』. 구미무역출판부, 1993.

[최기선, 2002] 최기선, 분야간 유사도와 통계기법을 이용한 전문용어의 자동 추출, 정보과학회논문지:소프트웨어 및 응용, 제29권 제3.4호, 2002년 4월, 2002.

[최수진, 2001] 최수진, 시소러스를 이용한 암묵지 저장 검색모형에 관한 연구, 경영논총 제19권 제2호, 2001.

[한국비서협회, 2000] 한국비서협회, 사무관리전문직 직무 분석, 창립1주년기념보고서, 2000.

[Berchtold, 2000] Berchtold, S., "Independent Quantification: An Index Compression Technique for High-Dimensional Data Spaces," Proceedings of ICDE Conference, Seattle, U.S.A. 2000.

[Crouth, 1990] Crouth, C., "An Approach to the Automatic Construction of Global Thesauri," *Information Processing and Management.*, 16(4), 1990.

[Davis, 1996] Davis, J. R., "The Networked Computer Science Technical Report Library," *IEEE Computer Special Issue on Building Large-scale Digital Libraries,* 1996.

[Gurda, 1995] Gurda, B., D. Danielsen, and B. Hemstead, "What Metadata Is and Why It Is Important," (http://badger.state.wi.us/agencies/wlib/sco/pages/qu-meta.htm), 1995.

[Khoshafian, 1997] Khoshafian, S., Multimedia and Imaging Databases, Morgan Kaufmann, 1997.

[Ricardo, 1999] Ricardo, B., *Modern Information Retrieval,* Addison Wesley, 1999.

[Sacks-Davis, 1992] Sacks-Davis, R., "Advanced database systems for text retrieval," *Proceedings of the 3rd Australian Database Conferences,* 1-8, 1992.

[Schatz, 1997] Schatz, B., "Building Large-Scale Digital Librarites," *IEEE Computer*, Vol.29, No.5, 1997.

[Shelley, 1995] Shelley, E. P., "Metadata: Concepts and Models," *Proceedings of the third National Conference on the Management of Geoscience Information and Data*, 1995.

[Subrahmanian, 2000] Subrahmanian, M., *Multimedia Database Systems*, Springer-Verlag, New York, 2000.

[Thanos, 1999] Thanos, C., *Multimedia Office Filing*, Amsterdam: North-Holland, 1999.

[Toshio, 1995] Toshio, Y., "The EDR Electronic Dictionary," *Communications of the ACM* 38(11): 42-44, 1995.

[Yates, 1996] Yates, B., "Integrating contents and structure in text retrieval, *ACM SIGMOD Rec.* 25, 1 (Mar.), 67-79, 1996.

[부록]

(국제회의관련 용어집 일부)

회의준비
- 회의일정
- 회의안내
  - 회의장소안내
  - 교통안내
  - 관광문화행사안내
- 회의일정표
  - 팜플렛
  - 브로슈어
- 회의비품
  - 명찰
  - 명패
  - 기념품
  - 필기구
  - 회의기기
  - 화상회의
  - PC
  - OHP
  - LCD프로젝터
  - 슬라이드
  - 화이트보드
  - 차트
  - 마이크
  - 녹음장치
  - 의제작성
  - 의제통지서

회의자료
- 회의자료다운로드
- 회의자료작성
- 회의자료배부

회의진행
- 등록접수
- 회의용어

회의참석자
- 참가자인선
  - 게스트
  - 의장
  - 사회자
  - 강사
- 참석자통지
- 출석확인

회의접대

회의장
- 회의장구성
- 회의장좌석안내

회의록

회의후정리
- 감사장
- 회의결과보고서
- 회의홍보
- 회의참가
- 내부회의출석
- 외부회의출석
- 파티초대
- 출결회답

국제매너
- 복장
- 인사예절
- 대화법
- 식사예절
- 전화예절

파티
- 조찬
- 련천
- 디너
- 다과회
- 티파티
- 칵테일파티
- 뷔페파티
- 가든파티
- 생일파티
- 축하파티
- 환영파티
- 환송파티

홍보
- 홍보업체
- 언론매체
- 홍보매체
- 안내장발송

행사준비
- 초청장발송

행사운영

행사후처리

해외출장
- 여권준비
  - 여권신청서류
- 비자신청

비자신청서류
환전
검역
여행수속
국가정보
국가공휴일
출장준비물
서류준비
선물준비
출장휴대품
출장중업무대행
출장중업무연락
수행출장
여정관리
출장일정표
출장보고서
여비
여비청구
여비정산
교통편예약
비행기예약
열차예약
자동차예약
숙박예약
숙박할인
방문처연락
문서수신
문서접수
정기적서류책자수신
문서접수대장
친전편지
수신물분류
수신물제출
출석여부답례
문서발신
발송인
문서발송대장
사내문서수발신
팩스송수신
우편업무
등기
라벨
요금별납
수취인부담
요금후납
우편번호
우편요금
우편주문판매
국제우편서비스
우편할인율
엽서
봉투
서신접는방법
문서관리
결재
전결
대결
후결
전자결재
문서관리시스템
문서분류
가나다식분류
주제별분류
번호식분류
지역별분류
명칭별분류
프로젝트별분류
형식별분류
문서정리
주제결정
상호참조표시
직인
문서보존
문서대출
대출기록부
문서열람
보존기간
영구
준영구
문서작성
letterhead
salutation
초청서
승낙서
소개장
추천장
견적서
의뢰서
주문서
항의장

저자소개

김명옥 (myongkim@ewha.ac.kr)

Computer Science, Ph.D., Wayne State University, Michigan, U.S.A.

현재 이화여자대학교 경영대학 부교수

관심분야: 사무정보시스템, 전문가시스템,데이터베이스, Conflict Resolution