파일명의 의미 클러스터링에 의한 윈도우 시소러스 WTPM 설계와 구현

김만필*, 차홍준*

Design and Implementation of The Windows Thesaurus WTPM using Filename of Semantics Clustering

Man-pil Kim*, Hong-jun Tcha*

요 약

객체지향 프로그래밍 언어를 기반으로 윈도우 사용자의 컴퓨터 파일시스템에 기록된 파일의 의미를 분석한 후, 이를 사용자 편의를 위해 파일명의 의미를 시소러스로 클러스터링 하는 설계를 하고, 파일로 기록된 문자의 의미와 파일확장자를 기반으로, 데이터베이스를 구성하고 참조하여 사용자 작성 파일들을 시소러스의 의미 체계와 통제어로 클러스터링 하여, 윈도우시스템의 화면표시 되는 Icon 파일들을 자동으로 분리하고, 설정하여, Mashup 시각구조로 나타내어 주는 프로세스(WTPM)를 설계하고 구현을 하였다.

ABSTRACT

Analyze semantic of files recorded in the user's computer file system based on C++ program language which pursue modularization program and object-oriented programming language. And this refers to it, it design that clustering semantic of filename with thesaurus for user convenience. WTPM makes User Write Files into Cluster with thesaurus semantic structure and reserved words. WTPM process has designed for Icon file's display Mashup structure and implemented by automation algorithm of classification

Key Word : thesaurus, filter driver, mashup, clustering, file system

Ⅰ. 서 론

21 세기 시·공간을 초월하는 정보유통의 도구로서 정보화 사회를 이루게 한 컴퓨터는 1942년 뉴우만의 프로그램 내장방식[1]에 의해 완성된 디지털컴퓨터로부터 시작이었다. 그리고 디지털 컴퓨터를 운영하는 운영체계는 1983년 MS-DOS부터 시작하여, 1995년 윈도우95를 거쳐, 2001년의 WindowsXP로 변화되었다.

이 때, 윈도우 화면의 특성으로 다양한 상징화면(icon)을 구사할 수 있도록, 또 파일을 가시적으로 볼 수 있도록 편의성을 제공하고 있지만, 사용자의 무의식적인 이용으로부터 생성되는 바탕화면의 파일 Icon의 정렬과 운영관리에서는 윈도우 시스템의 기능에서 벗어난 파일 화면관리 문제가 나타났다.

* 강원대학교 컴퓨터학부(kimp67@kangwon.ac.kr)

접수일자 : 2009.01.16
완료일자 : 2009.02.20
접수번호 : KIIECT2009-01-10

그러므로 이 연구는 모듈화 프로그램을 추구하는 C++ 언어와 같은 객체지향 프로그래밍[2] 언어를 기반으로 윈도우 사용자의 컴퓨터 파일 시스템에 기록된 파일의 의미를 분석한 후, 이를 사용자 편의를 위해 파일명의 의미를 시소러스[3]로 클러스터링[4] 하는 설계를 하고, 이를 작성할 때는 이를 Mashup[5] 시각구조로 화면에 표시 되도록 목적하였다. 이를 이루기 위하여 윈도우 파일 시스템에서 파일로 기록된 문자의 의미와 파일확장자를 기반으로, 데이터베이스를 구성하고 참조하여 사용자 작성 파일들을 시소러스의 의미 체계와 통제어로 클러스터링 하여, 윈도우시스템의 화면표시 되는 Icon 파일들을 자동으로 분리하고, 설정하여, Mashup 시각구조로 나타내어 주는 프로세스를 설계하고 구현을 하였다.

II. 파일정보와 시소러스

이 장에서는 윈도우시스템에서의 파일시스템에 관한 구조와 운영기술과 시스템프로세서들의 정보를 구분하고, 분류할 수 있는 운영체제를 탐색하여, 그 파일명이 가지는 어휘를 인식할 수 있는 통제와 시소러스의 구조 작성을 문헌조사 한다.

2.1 파일정보의 구성과 색인

윈도우시스템에서의 파일(file)은 한 단위로 이루어지는 정보의 모음으로서 데이터 세트(data set)라 하지만, 정보처리에서 파일의 기본 단위는 같은 속성에 해당하는 데이터들의 항목인 필드(field)와 그 값이며, 서로 관련 있는 값들의 집합으로 레코드(record)라는 모임(group)을 장치에 저장할 수 있도록 가시적인 형태로 이룬 것이다.
그러므로, 이 파일시스템(file system)은 윈도우시스템에서 정보와 데이터를 실제로 운용하는 파일을 생성, 갱신, 검색, 관리할 수 있도록 체제를 갖추고 있으며, 그 파일의 내부에는 정보를 기록하는 구조적인 일과 이를 운영체제가 운영하면서 관리하도록 되어있다. 파일정보는 파일의 처리를 목적하여 시스템으로 운영되고 관리되어지는 파일명(file name)에 의해서 나타내어지는 속성을 의미한다. 파일명이 자동으로 문서분류가 되려면, 문서 제목으로부터 필요한 색인추출(indexing) 과정이 필요해 지기 때문이다.
파일명의 색인은 불용어 제거 알고리즘, 어근 추출 알고리즘, 동의어 사전, TF*IDF 알고리즘과 벡터 길이 정규화 과정을 거쳐 이루어진다.

2.2 파일정보의 어휘의 통제

파일정보의 어휘는 자연어와 통제어로 구분한다. 자연언어(自然語: Human Language)는 사람들의 모든 사회에서 의사소통의 도구로 사용되는 원천 언어로서, 이는 사람의 생각이나, 느낌, 소리, 혹은 글자로 나타내는 수단이다. 그러나 형식언어(型式言語)는 명제(命題)나 법칙과 같이 명확한 수식에 의해서 설명할 수는 없지만, 부정할 수도 없이 긍정적으로 인정하여야 할 논리적 시스템에 관계 되어지는 형식시스템(formal system)으로 설계하고, 이를 구현하려는 프로그래밍 언어이다. 이 같은 형식언어(ξ)는 6가지의 집합요소(tuple)의 규정을 갖추어야 만이 정의되어 진다[1].

(정의) $\xi = (C, V, M, P, S, D)$

C: 유한 규정(canons)
V: ξ에 의해 입증할 수 있는 자모 열(alphabet)
M: 유한 문자나 기호의 변수(variable)
P: 유한 기호로 기술된 술부(predicate)
S: 유한의 구두(句讀: punctuation)
D: 술부에 속한 술부문장(sentence)

그림 1. 형식언어의 6가지 요소
Fig. 1. six tuple of formal language

어휘의 통제는 특정 분야의 문서를 작성할 때 지켜야하는 원칙들과 제약 조건들을 명시해 주는 자연언어 체계를 인위적으로 통제하여 얻게 하려는 것이다. 그러므로 통제언어의 개발은 내용의 명확한 전달과 이해가 정확한 공정을 위해 필수적이어서, '어떻게 가독성을 높일 것인가?'와 ' 어떻게 번역(기계번역을 포함해서)을 할 것인가?' 하는 것으로 요약될 수 있다. 따라서 어휘의 통제로서 가장 좋은 형식언어는 메타언어가 있다.

즉, 메타언어(meta language)는 어떤 주제(主題: subject)에 대응하여 주체(主體)의 물질적 활동이나, 인식활동을 지시하는 방향의 대상(對象: object)을 표현하는 대상언어가 아닌, 그 대상언어에 관하여 말하려는 언어이다[2].

2.3 시소러스의 구조와 작성법

특정주제 분야의 용어의 그룹화와 그 구성을 결정할 때에 이용되는 구분방법은 다양하지만, 이용 가능한 분류방법에는 '하나의 방법에 따른 클래스에 의한 배열'이라는 의미에서 체계적인 분류, 혹은 자동분류가 된다. 즉, 이는 체계적인 분류법에서 일련의 클래스를 모아, 이를 코딩시스템인 기호법을 이용해 구성하는 것이며, 또 체계적인 분류의 일종으로 Passing분류로 기본적인 카테고리에 대응 시켜서 배열하는 분류방법이 되기 때문이다. 분류의 관계에는 등가관계, 계층관계, 연관관계를 들 수 있다.

등가관계란 색인 작업 시 복수의 용어가 동일개념을 나타내고 있다고 인정되는 경우에 우선 어(語) 및 비 우선어간의 관계이다. 즉, 우선 어는 그 개념을 표현하기 위하여 색인 작업 시에 이용되는 용어를 말하며, 비우선 어는 이용되지 않는 용어로서, 이들 어는 등가어의 집합으로 형성된다.

계층관계는 상위와 하위의 수준을 나타내기 위하여, 상위어를 클래스, 또는 전체로 나타내며, 하위어는 그 한 요소, 또는 일부분으로 나타내는 것이다. 즉, 계층관계는 상위와 하위 개념을 논리적으로 전개하는 순서로 위치시키기 위하여, 이용되는 것으로 시소러스와 구조화되어 있지 않는 용어리스트와 구별하는 기본적인 특징을 재현성으로 제고시키며, 적합성도 제고시키는 중요한 요소가 된다.

연관관계란 계층적이 아니고, 개념적으로는 밀접하게 관련되어 있으나, 등가집합에는 포함되지 않는 용어간의 관계이다. 즉, 표준규칙에 연관관계를 가지는 색인 작성과 탐색에 이용되는 대체용어로 용어간의 연결을 명시하는 것이다. 그러므로 연관관계는 시소러스 편찬자가 기술(記述)하는 것으로, 그 결과 재현성과 적합성 향상을 저하시킬 우려가 있는 문제에 관련하게 된다.

실제로, 연관관계 어는 색인을 이용하는 사람의 사고방식과 범위에 근거하여, 한쪽의 용어가 색인어로 채택될 때, 다른 쪽의 용어로 강력하게 암시되어지는 것으로, 이는 한쪽의 용어가 다른 쪽의 용어를 정의하거나, 설명하기 위하여 필요한 구성요소를 가지게 되는 것이다.

시소러스의 작성은 파일명의 주제를 명확히 분야를 설정하도록, 그 주제의 경계를 확정해 다루어야 할 부분과 주변을 구별하여야 한다. 그리고 편찬자는 시스템이 요구하는 사항에서 전체를 조사하여, 어떠한 종류의 시소러스를 작성할 지를 명확히 해야 한다. 또한 시소러스를 연역적으로 편찬하려는 경우에는 충분한 수의 용어를 수집한 후에 선택한 용어로 검토하며, 또 그 구조를 발견하고 어휘를 통제하게 된다. 그러나 귀납적 방법으로 편찬하려는 경우에는 용어가 파일상에 출현함과 동시에 시소러스에 추가하여, 색인 작업에서도 사용할 수 있게 하므로, 그 어휘통제를 처음부터 실시하게 할 수 있도록, 그 이상의 상위 카테고리에서도 부여하는 것이다.

그러므로 색인 작업은 작성 초기부터 실시하나, 사용하던 용어의 의미가 나중에라도 확실해지면, 그 색인 작성을 수정할 수 있으므로 특정의 어휘통제는 초기단계부터 적용하는 것이 좋다.

Ⅲ. 윈도우 파일의 구성

3.1 윈도우의 구조

윈도우는 사용자가 이용의 편의성(friendly)으로, 그래픽스로 표현되는 화면의 크기와 모양을 마음대로 조절할 수 있을 뿐만이 아니라, 다중처리(multitasking) 시스템으로 윈도우 화면을 여러 개로 중첩해 띄울 수 있도록, 또 네트웍 개방형 플랫폼(platform)으로 개발한 사용자 워크스테이션(workstation) 컴퓨터 안에 속해 있는 프로그램(client)들을 서비스로 요청할 수 있게 그림 2.에서와 같은 구조를 갖추고 있다. 즉, 이 Windows 구조는 어플리케이션이 운영되는 User Mode에서 Windows Display로 동작하는 다른 프로그램들처럼 특권 없이 실행되는 프로세서 모드와 커널, 메모리 관리자, 캐시 관리자, 프로세스 관리자, 프로세스간 통신을 위한 LPC와 RPC, 오브젝트 관리자, I/O 관리자, 시스템의 환경 설정 관리자 HAL (Hardware Abstraction Layer)등이 포함되어서 실행하는 Kernel Mode로 나누어진다.

3.2 윈도우 파일시스템의 구조

Windows 운영체제의 계층 구조는 디바이스 스택(device stack)이 위치하는 영역에 따라 커널 모드 드라이버 계층상 Higt Level로 FAT, NTFS, CDF와 같은 파일 시스템과 Low Level 드라이버는 버스에 장착된 USB, PCI, ISA, SCSI, IEEE 1394, 등을 제어하는 드라이버로서, 이는 직접 하드웨어 장치들을 제어할 수 있는 Intermediate를 지원하는 드라이버 파일 시스템을 가진다. 윈도우의 파일 시스템 드라이버(File System Driver)는 보조 저장장치의 데이터를 관리하기 위하여 하드웨어적으로 파일을 저장하기 위한 것이다. 즉, 이는 저장장치를 관리하는 서브 시스템의 구성 요소로서, 드라이버의 계층 구조상 하드웨어 장치와 별도로 독립적인 I/O 요청을 처리하는 드라이버로 그림 3에서와 같이 디스크와 같은 보조 저장 장치에 저장된 데이터를 가져와 서로 연관된 데이터를 하나의 폴더나 파일 단위로 된 객체로 구성하여, 사용자 입장에서 데이터 관리의 편의를 제공하게 한다.

또한 로컬 파일 시스템 드라이버는 논리적인 볼륨 관리자(Logical Volumn Manager)는 파일 시스템 드라이버가 접근할 수 있는 디스크 공간을 알려주고 논리적으로 구성된 블록을 실제 물리적으로 구성된 Block과 Mapping을 하게 한다. 필터 드라이버(Filter Driver)는 일종의 Intermediate 드라이버이다. 이는 파일시스템 드라이버나 디스크 드라이버와 같이 상용화되어 있는 드라이버에 전달하는 I/O 요청을 가로채어 기존의 드라이버가제공하는 기능을 보완하거나 새로운 기능을 추가 할 수 있는 기회를 제공한다.

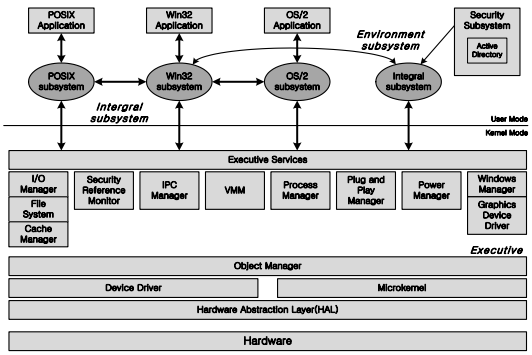

그림 2. Windows 시스템의 구조
Fig. 2. Construction of Windows System

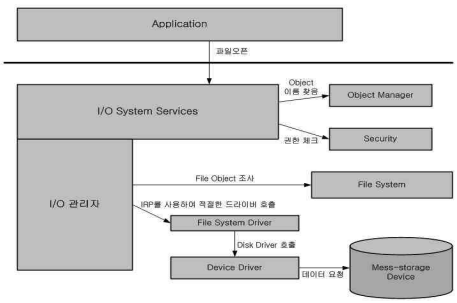

그림 3. I/O 관리자와 오브젝트 관리자
Fig. 3. I/O Manager and Object Manager

그림 4. 파일 시스템 필터 드라이버
Fig. 4. Filter Driver of File System

그림 4.에서와 같이 드라이버 계층에서 드라이버에 전달되는 I/O 요청을 어디서 가로채느냐에 따라, 상위(Upper)필터 드라이버와 하위(Lower)필터 드라이버로 나누게 된다. 상위 필터 드라이버는 사용자 프로세스의 요청이 파일 시스템 드라이버에 도달하기 전에 I/O 요청을 가로채어 필요한 작업을 하고, 하위 필터 드라이버는 파일 시스템 드라이버 아래에 위치한 파일 시스템 드라이버에서 처리한 I/O 요청을 하드디스크와 같은 보조 저장 장치로 관리하는 드라이버에 전달할 때, 이 I/O요청을 가로채어 필요한 작업을 하게 된다.

Ⅳ. 윈도우시소러스 WTPM의 설계

4.1 시소러스의 요소와 요구

시소러스를 구성하기 위한 관점들로서 요소는 시스템의 부분으로 파일들 간의 관련성 정보를 이용하여 파일을 집단화하고, 색인의 질을 높이며, 검색 효율을 향상시키려는 것이다. 그러므로 문서를 대표하는 단어에서 파일명으로 사용자의 의도를 파악하는 사용자 속성파일(user profile)로 분류(classification)하고, 여과(filtering)하여야 한다. 윈도우시스템에서 파일명의 어휘는 컴퓨터 자료처리에 사용된 사용자파일들의 이름으로 자연언어와 통제언어를 혼합한 시스템을 구축하기 위하여, 동의어의 어근 추출을 위한 데이터베이스(DB) 사전을 만들기 위한 것이며, 파일의 색인이 되는 기반이다. 표 1.에서와 같이 시소러스 클러스터링으로 어휘를 통제할 동의어 사전 형식으로 설정한다. 따라서 표 1.의 형식에 시스템 파일, 라이브러리 파일, 패키지 파일은 시스템 설치, 혹은 각종 드라이브를 설치할 때 등록되어 지는 모든 파일들이며, 사용자 파일로 구분하게 되는 데, 이 때

표 1 시소러스 DB파일 형식구조
Table. 1 DB File Format of Thesaurus

시소러스구분	파일명 어휘	Icon	Icon 색인
시스템 파일	- - -	* @ # -	Icons01 Icons02 - -
라이브러리 파일	- -		Iconl01 Iconl02 - -
팩키이지 파일			Iconp01
사용자 파일			Iconu01

파일명은 윈도우시스템의 파일시스템으로부터 검색(dump) 받아 작성하게 된다. 파일의 불용어로는 컴퓨터 분야의 경우 “computer”, “program”, “source”, “machine”, 그리고 “language”등과 같이 자주 사용되어지는 Reserved word인 단어와 어휘를 불용어로 처리한다. 또한 파일명의 길이는 TF*IDF 알고리즘을 통하여 15자 이내로 만든다.

이 같은 요소를 만족시키기 위하여 파일 가로채기(hooking), 시소러스의 작성, Icon DB, WtPM 드라이버의 구성과 설치가 요구 된다.

WtPM 드라이버의 구성은 그림 5.에서와 같이 Window I/O System에서 User Mode와 Kernel Mode로 구분된 윈도우시스템 운영체제가 실행되어질 때, 모든 시스템 윈도우 입출력 정보가 나타나게 되는 데, 이 때 IRP_MJ_WRITE에 의해서 File System Driver에 상주되어 있는 각각의 Device Driver가 I/O Stack Location에 상호참

조 하면서 정보를 교환하게 된다. 여기서, 윈도우시소러스 실행 모듈인 WTPM이 실행하여 윈도우 창에 나타내어야 할 파일명 정보를 WtPM_filter_Driver가 IRP와 FSD(File System Driver) 사이에서 윈도우 파일명을 가로채기(file Hooking) 하도록 설치하는 것이다.

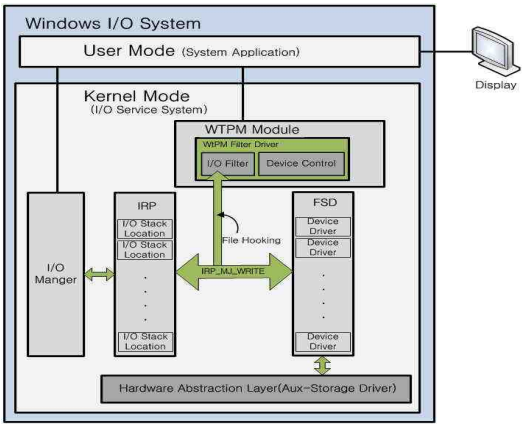

그림 5. Window I/O System의 WtPM설치 구조
Fig. 5. Construction WtPM of Window I/O System

4.2 WTPM 프로시저

윈도우시소러스 표현 프로시저 모듈(WTPM)은 데이터베이스(DB)와 4개의 프로세서를 가진 프로시저로 그림 6에서와 같은 구조로 구성된다. WTPM을 위한 데이터베이스(DB)는 윈도우시소러스 파일의 DB, 불용어 파일의 DB, 그리고 WTPM의 DB로 구성된다. 이는 WTPM이 윈도우 시스템 운영체제인 BIOS와 커널 드라이브에 의해서 시스템 파일을 참조 받아 WTPM의 프로시저가 활용하게 된다. WTPM의 프로시저로서 이

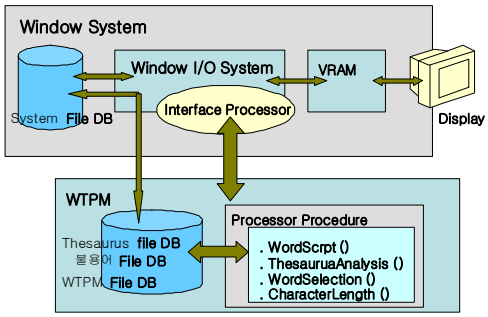

그림 6. WTPM 시스템 구조
Fig. 6. Construction of WTPM

는 시스템파일의 DB로부터 파일명을 가로채기하여 파일명 어휘를 결정하는 프로세서 “WordScrpt()"와 윈도우시소러스를 분석해내는 프로세서 "ThesaurusAnanlysis()", 불용어를 처리하여 파일이름을 설정해 주는 프로세서 ”WordSelection()“, 그리고 파일명의 길이를 확정하면서 Icon과 등치 시키는 프로세서 ”CharacterLength()“로 구성된다.

V. 결 론

이 연구는 윈도우 파일 시스템에서 시스템 및 사용자 작성 파일들로부터 Mashup 시각구조로 나타내어 주는 WTPM 시스템 구조를 설계하고 구현하였다. 이 과정에서 문헌정보의 용어 기록의 형식에 따라 시소러스의 의미 체계와 통제어로 클러스터링 하여, 표 1. 시소러스 DB파일 형식구조를 설계하고, 이를 구현하였고, Window I/O System에 WtPM 드라이버를 설치하여 윈도우즈 바탕화면의 윈도우즈시스템 운영체제의 커널에 WtPM 드라이버를 저장하고, 이것이 실시간 모니터링을 하여 추출한 파일명이 시소러스 클러스터링 Icon으로 Display된다는 평가를 받았다. 그러나 윈도우시스템의 모든 파일에서 완전한 의미파일 시소러스 클러스터링 처리가 될 수 있어야 할 Data Meaning에 의한 지능화 DB 문제가 해결과제로 남겨졌다.

참 고 문 헌

[1] John von Neumann; First Draft of a Report on the EDVAC, February 1945.

[2] Wegmann A.; "Object-Oriented Programming Using Modula-2.",Journal of Pascal, Ada ' Modula-2, pp.43-51, Vol.5 No.3, 1986.

[3] 김태수; “용어 정의를 도입한 시소러스 개발 연구”,정보관리학회지,pp.231-254. Vol.18 No.2, 2001.

[4] Silverman, Ellen-Marie; “Clustering". The Journal of speech and hearing disorders, pp.24-32, Vol.16 No.4, 1973.

[5] Miller, C. C.; "A Beast in the Field: The Google Maps Mashup as GIS/2" Cartographica, pp.6-15, Vol.41 No.3, 2006.

[6] 김기영; 대명사의 본질-자연언어와 형식언어의 차이, 한국독어독문학회, 2003.

[7] Delebecque, H.; "HDSML: an Hyper Document Structuring Meta Language", ED MEDIA-PROCEEDINGS, Vol.2 No.2, 2004.

저자약력

김 만 필(Man-Pil Kim)

2000년 강원대학교 전자계산학과 학사
2003년 강원대학교 컴퓨터과학과 석사
2008년 강원대학교 컴퓨터과학과 박사

<관심분야> 시스템프로그래밍, 이미지프로세싱, 정보보안, GIS

차 홍 준(Hong-Jun Tcha)

1991년 성균관대학교 통계학과 박사
2004년 노동부 기술사검정위원
현재 강원대학교 컴퓨터과학과 교수

<관심분야> 시스템프로그래밍, GIS, 전산통계