인터넷 자원의 표현·전송을 위한 구문

미국 콜로라도 대학에서 개발한 인터넷 종합정보시스템인 'Harvest'는 인터넷 자원을 표현하고 이를 전송하기 위한 기계가독형 구문으로서 SOIF를 사용한다. 객체 개념을 도입한 SOIF는 각각의 자원에 대한 기술을 객체로 저장하고 하나의 SOIF 스트림에 여러 개의 객체를 동시에 전송하게 된다. SOIF는 인터넷상의 각종 자원을 표현하고 이를 네트워크상에서 운용할 수 있는 효과적인 메타데이터이다.

이재진/ 한국데이터베이스진흥센터 정책연구과

연재 순서

1. 개요

1994년 1월 Harvest 프로젝트의 일환으로 처음 정의된 SOIF(Summary Object Interchange Format)는 파일, 사이트, 웹 페이지와 같은 인터넷 자원을 구조화시킨 요약 객체(summary object)로 생성·기술하고 이를 전송하기 위한 기계가독형 구문이다. SOIF는 IAFA 템플릿과 BibTeX 서지 형식에 기초하여 설계되었고 스트림내에서 객체를 구분하고 객체에 대한 접근과 호출이 용이하도록 URL을 별도로 관리하는 특징을 갖는다. 이 때 객체는 개별 자원을 의미한다.

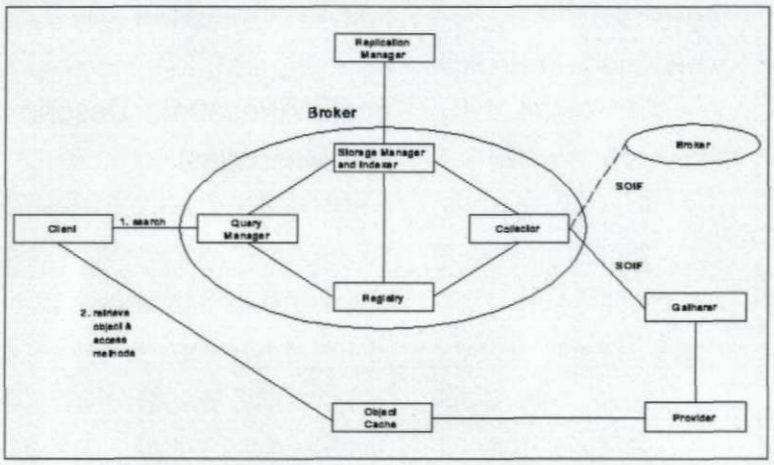
Harvest는 인터넷상에 있는 적합한 정보의 수집, 추출, 조직, 탐색, 저장, 복제를 위한 종합적인 도구로서 이 시스템의 가장 큰 장점은 SOIF 요약 색인을 구축하기 위한 데이터 수집 아키텍처를 제공한다는 것이다. 이는 이용자가 직접 색인 템플릿을 작성해야 하는 WHOIS++이나 색인 데이터의 수집 방법이 정의되어 있지 않은 GILS와 대조되는 측면으로서 이용자가 직접 템플릿을 이용하여 색인을 구축하거나 시스템이 자동으로 생성할 수 있다.

SOIF는 Harvest에서 사용하기 위해 개발되었지만 인터넷 자원 탐색 메카니즘인 RDM (Resource Description Messages)과 미국 스탠포드 대학의 STARTS 프로젝트 등에서 활용되고 있다.

2. Harvest 시스템

Harvest는 여러 개의 하부시스템으로 구성된다. Gatherer하부시스템은 Provider사이트(FTP나 HTTP 서버)로부터 제공이 가능한 자원에서 색인 정보(키워드, 저자명, 표제 등)를 수집한다. Broker 하부시스템은 하나 이상의 Ga-therer에서 색인 정보를 검색하고 중복된 정보를 제거하여 수집된 정보를 색인한 뒤 WWW 질의 인터페이스를 제공한다. Replicator하부시스템은 인터넷에 Broker를 효과적으로 복제한다. 이용자는 Cache하부시스템을 통해 정보를 효율적으로 검색할 수 있다. Harvest Server Registry(HSR)는 인터넷상의 Gatherer, Broker, Cache, Replicator에 대한 각각의 정보를 가지고 있어 새로 Gatherer나 Broker를 생성하거나 적절한 Broker를 찾을 때 유용하게 이용된다.

이 때 Gatherer와 Broker는 속성-값 쌍의 스트림 프로토콜인 SOIF를 사용하여 서로 메시지를 교환한다. 즉 Gatherer는 SOIF에 개별 객체에 대한 내용 요약을 생성하고 이 요약을 수집하고 색인하는 Broker에게 전달하는 것이다. SOIF는 요약 객체를 포괄하여 다루면서도 동시에 Broker가 Gatherer로부터 SOIF 내용 요



〈그림〉 Harvest 아키텍처

약을 검색하도록 한다. Broker는 구조화된 속성-값 질의와 여러가지 서로 다른 유형의 질의를 사용하여 SOIF 데이터에 대해 질의할 수 있도록 한다.

3. 기본 구성요소

각각의 SOIF 객체는 템플릿 유형(template type), URL, 0개 이상의 속성-값 쌍(attribute- value pairs)의 3가지 기본 요소로 구성된다.

템플릿 유형은 특정 SOIF 객체에 포함된 속성 집합을 식별하는데 사용된다. SOIF는 템플릿 유형 자체를 정의하지 않고 이미 정의된 템플릿 유형명과 요약 객체를 연결하는 방법을 제공한다. 이러한 템플릿 유형은 IAFA 템플릿 유형에서 유래된 것으로, Harvest 시스템에서 생성되는 SOIF 객체는 'FILE' 템플릿 유형을 갖는다. 'FILE' 템플릿 유형은 다양하고 방대한 웹기반 자원을 가리키는 일반적인 템플릿 유형이다. 이밖에도 'DOCUMENT'와 'OBJECT' 등의 템플릿 유형이 있다.

URL은 SOIF의 객체 IDENTIFIER로 사용된다. SOIF는 URL을 사용하여 접근할 수 있는 네트워크 자원과 개별 요약 객체를 연결한다. 이러한 연결은 같은 자원이 URL로 구별되는 다중 요약 객체를 갖도록 한다.

속성-값 쌍은 URL로 참조되는 네트워크 관련 자원에 대한 메타데이터를 포함한다. 이들 속성-값 쌍은 속성 식별자(IDENTIFIER), 값의 길이, 구분기호, 그리고 값으로 구성된다. SOIF 구문은 하나의 속성이 여러 개의 값을 가지는 것을 금하며 만일 같은 속성이 여러 개의 값을 가져야 하는 경우에는 속성명에 하이픈(-)과 양의 정수를 추가하여 속성 식별기호를 생성한다. 예를 들어 3명의 저자를 갖는 특정 자원이 있다면 속성 'Author'는 속성 식별자로 'Author-1', 'Author-2', 'Author-3'를 사용할 수 있다.

4. 공통 SOIF 속성

Harvest 시스템에서는 각 Broker가 소장하고 있는 데이터에 따라 서로 다른 속성을 지원할 수 있다. 〈표 1〉은 여러 속성 가운데 공통적인 SOIF 속성을 나타낸 것이다.

〈표 1〉에서 진하게 표시된 초록(Abstract), 저자(Author), 기술(Description), 키워드(Keyword), 서명(Title)은 기본적인 기술 요소이며 유형(Type)의 사례는

〈표 1〉 SOIF 공통 속성

속 성	기 술 (내용)
Abstract	객체에 대한 간단한 초록
Author	객체의 저자
Description	객체에 대한 간단한 기술
File-Size	객체의 크기(바이트수)
Full-Text	객체의 전체 내용
Gatherer-Host	Gatherer가 객체에서 정보를 추출해 내는 호스트
Gatherer-Name	객체에서 정보를 추출한 Gatherer 이름
Gatherer-Port	Gatherer의 정보를 제공하는 Gatherer-Host의 포트번호
Gatherer-Version	Gatherer의 버전번호
Keywords	객체에서 추출한 탐색가능한 키워드
Last-Modification-Time	객체의 최종 수정 시간
MD5	객체의 MD5 16바이트 누계
Refresh-Rate	최종 수정 시점과 요약 객체가 재생성되는 시간 차이 초기값 1개월
Time-to-Live	최종 수정 시점으로부터 요약 객체의 유효기간 초기값 6개월
Title	객체의 표제
Type	객체의 유형. 〈사례〉 Archive, Audio, Awk, Backup, Binary, C, CHeader, Command, Compressed, CompressedTar, Configuration, Data, Directory, DotFile, Dvi, FAQ, FYI, Font, FormattedText, GDBM, GNUCompressed, GNUCompressedTar, HTML, Image, Internet-Draft, MacCompressed, Mail, Makefile, ManPage, Object, OtherCode, PCCompressed, Patch, Perl, PostScript, RCS, README, RFC, SCCS, ShellArchive, Tar, Tcl, Tex, Text, Troff, Uuencoded, WaisSource
Update-Time	Gatherer가 객체의 내용 요약을 갱신한 시간
URL-References	HTML 객체에 있는 모든 URL 참조

```
SOIF            ::= OBJECT SOIF | OBJECT
OBJECT          ::= @ TEMPLATE-TYPE { URL ATTRIBUTE-LIST }
TEMPLATE-TYPE   ::= IDENTIFIER
ATTRIBUTE-LIST  ::= ATTRIBUTE ATTRIBUTE-LIST | ATTRIBUTE |
                    NULL
ATTRIBUTE       ::= IDENTIFIER (VALUE-SIZE) DELIMITER VALUE
URL             ::= RFC1738-URL-Syntax | "-"
IDENTIFIER      ::= ALPHA-NUMERIC-STRING
VALUE           ::= ARBITRARY-DATA
VALUE-SIZE      ::= NUMERIC-STRING
DELIMITER       ::= ":(TAB)"
```

〈예 1〉 SOIF 구문 문법

```
@DOCUMENT { http://home.netscape.com:80/
Title{19}: Welcome to Netscape
Content-Type{9}: text/html
Content-Length{5}: 33262
}

@DOCUMENT { http://home.netscape.com/eng/ssl3/ssl-toc.html
Title{19}: SSL Protocol V. 3.0
Content-Type{9}: text/html
Content-Length{5}: 5870
Author-1{14}: Alan O. Freier
Author-2{14}: Philip Karlton
Author-3{14}: Paul C. Kocher
Abstract{318}: This document specifies Version 3.0 of the (B)Secure
Sockets Layer (SSL V3.0)(/B) protocol, a security protocol that
provides communications privacy over the Internet. The protocol allows
client/server applications to communicate in a way that is designed
to prevent eavesdropping, tampering, or message forgery.
}
```

〈예 2〉 'DOCUMENT' 템플릿 유형을 사용한 SOIF 요약 객체

대부분 다양한 컴퓨터 파일 형식이나 언어와 관련된 것으로 SOIF에서 네트워크를 통하여 접근 가능한 객체를 색인하기 위한 것이다. 이밖에 필요한 속성은 사용기관에서 임의로 추가하여 사용할 수 있다.

5. SOIF 구문

SOIF 구문은 〈예 1〉과 같은 문법으로 정의된다.

URL은 RFC1738에서 정의된 구문을 사용하며 관련된 URL이 없을 경우, 하이픈(-)으로 대체한다. 식별기호(IDENTIFIER)는 알파벳 문자와 숫자, 하이픈(-)이나 밑줄(_)만을 사용한다. 값(VALUE)은 템플릿 유형에 근거하여 수신자가 인식할 수 있는 임의의 포맷으로 된 데이터를 포함하게 된다.

값크기(VALUE-SIZE)는 양의 정수로서 구분기호 뒤에 값이 차지하게 되는 옥텟(octet) 수를 가리킨다. 구분기호(DELIMITER)는 콜론(:)과 탭(\t)으로 구성된 두자리 옥텟 구분기호이다. { }는 URL과 속성리스트(ATTRIBUTE-LIST)의 조합뿐만 아니라 값크기를 묶어 표시하는데 사용한다.

@ 템플릿 유형(TEMPLATE-TYPE)은 SOIF 객체의 시작을 나타낸다. 숫자열(NUMERIC-STRING)은 0이나 0이상의 ASCII 숫자를, 알파벳-숫자열(ALPHA- NUMERIC-STRING)은 0이나 0이상의 ASCII 문자나 숫자, 하이픈과 밑줄을 나타낸다.

"DOCUMENT" 템플릿 유형을 사용한 SOIF 요약 객체는 〈예 2〉와 같다.

〈표 2〉 RDM의 요청/응답 모델

RD Retrieval	요청자가 서버에서 하나이상의 RD를 검색
RD Submission	요청자가 하나이상의 RD를 보내고 갱신 · 삭제를 위해 서버에 선택적인 스키마 정의를 보냄
Server Description Retrieval	요청자가 서버로부터 하나의 서버 기술을 검색
Schema Description Retrieval	요청자가 서버로부터 하나의 스키마 정의를 검색
Taxonomy Description Retrieval	요청자가 서버로부터 하나의 분류 정의를 검색
Status Retrieval	요청자가 서버로부터 현 서버 상태를 검색

6. RDM(Resource Description Messages)

RDM은 자원 기술(RD : Resource Description)로 알려져 있는 네트워크 자원에 대한 메타데이터를 탐색하고 검색하기 위한 메카니즘으로서 두 개의 프로세스가 네트워크를 통해 자원 기술을 교환할 수 있는 메시지 포맷이다. RDM에서는 하나의 프로세스(클라이언트나 에이전트)가 다른 프로세스(서버)에 RDM 요청 메시지를 보내면 RDM 응답 메시지를 받게 된다. RDM은 일정 범위내의 자원 기술을 한꺼번에 전송하거나 목록 개념을 지원하여 하나의 RDM 서버가 여러 목록에 접근할 수 있도록 한다.

RDM은 자원 기술을 코드화하기 위해 Harvest 시스템의 SOIF 포맷을 사용한다. SOIF가 제공하는 데이터 모델은 속성에 대해 평면적으로 기술하면서도 해당 값을 모두 객체인 BLOB(Binary Large OBjects)으로 처리한다. 자원 기술은 속성-값 쌍의 리스트로 구성되며 URL을 통해 자원간의 연결이 이루어진다. 에이전트가 로봇 등을 이용하여 자원 기술을 자동으로 생성할 수도 있고, 사서나 저자가 이를 수작업으로 직접 작성할 수도 있다.

RDM의 요청/응답 모델은 〈표 2〉와 같이 구성된다.

각각의 RDM 메시지는 헤더(header)와 본문(body)으로 구성된다. 헤더는 RDM 메시지의 특성과 적용되는 목록의 특징을 식별하며 본문은 필요한 요청을 수행하는데 요구되는 데이터를 포함한다. RDM 메시지의 헤더와 본문은 SOIF를 사용하여 코딩된다.

RDM은 1996년 넷스케이프의 목록 서버로 사용하고 이를 W3C에서 표준으로 제정하려는 움직임이 있었으나 현재는 RDF 등으로 대체되고 있는 실정이다.

7. 메타데이터 기술 요소의 비교

SOIF에서 사용되는 공통 속성 요소와 DC, IAFA 템플릿간의 상호 변환이 가능한 요소를 비교해 보면 〈표 3〉과 같다. 〈표 3〉에서 보는 바와 같이 SOIF는 속성 요소에 객체, 템플릿을 관리하기 위한 속성을 상대적으로 많이 포함하고 있다. 이는 SOIF가 인터넷상의 자원을 객체로 표현하여 이를 네트워크상에서 전송하기 위한 구문으로 사용되기 때문이다.

〈표 3〉 SOIF, DC, IAFA 기술요소 비교

SOIF	Dublic Core	IAFA 템플릿
abstract	DC.Description	–
author	DC.Creator	–
description	DC.Description	Description
file-size	–	Size-v1
full-text	–	–
gatherer-host	–	–
gatherer-name	–	–
gatherer-port	–	–
gatherer-version	–	–
keywords	DC.Subject	Keywords
last-modification-time	DC.Date.Created/ Modification_of_ present_form	Last-Revision-Date-v1
md5	–	–
refresh-rate	–	–
time-to-live	–	–
title	DC.Title	Title
type	DC.Type	Format-v1
update-time	DC.Date	–
url-references	DC.Relation	–

8. SGML, HTML 문서의 SOIF로의 변환

SGML 데이터를 SOIF로 매핑하는 데에는 4가지 방법이 있다. 처음 두가지는 SGML 태그의 내용(content)을 SOIF 속성으로 변환하는 것이며 나머지 두가지는 SGML 속성의 값(value)을 SOIF 속성으로 변환하는 것이다. 〈표 4〉는 변환 방법에 대한 설명과 각각의 예를 나타낸 것이다.

HTML은 SGML의 하부집합으로서 웹 문서의 기본적인 형식으로 널리 사용되고 있다. HTML의 각 요소는 〈표 5〉와 같이 SOIF 속성으로 매핑될 수 있다.

예를 들어 HTML에서 문서 표제가 '〈TITLE〉My Home Page〈/TITLE〉'인 경우, 이를 SOIF로 나타내면 'title{13}: My Home Page'로 변형되며 HTML 문서에서 URL 참조를 '〈A HREF="http://harvest.cs.colorado. edu/"〉'로 나타낸 경우, 이를 SOIF로 변형하면 'urlre-ferences{32}: http:// harvest.cs.colorado. edu/'가 된다.

〈표 4〉 SGML 데이터의 SOIF로의 변환 방법

변환 유형	SGML	SOIF	내 용
내용(content)	〈TAG〉	soif1, soif2, …	태그 "TAG"를 SOIF 속성 "soif1", "soif2"로 변환
	〈TAG, ATT=x〉 〈TAG, ATT=y〉 〈TAG〉	x-stuff y-stuff stuff	"ATT"가 "TAG"의 속성이면 속성값에 따라 SOIF의 속성을 다르게 표시
값(value)	〈TAG:ATT〉	att-stuff	"TAG" 태그의 "ATT" 속성의 값을 "att-stuff"로 표시
	〈TAG:ATT1〉	$ATT2	SGML 속성의 값을 다른 SOIF 속성으로 표기

〈표 5〉 HTML 요소와 SOIF 속성

HTML 요소	SOIF 속성
〈A〉	keywords, parent
〈A:HREF〉	url-references
〈ADDRESS〉	address
〈B〉	keywords, parent
〈BODY〉	body
〈CITE〉	references
〈CODE〉	ignore
〈EM〉	keywords, parent
〈H1〉	headings
〈H2〉	headings
〈H3〉	headings
〈H4〉	headings
〈H5〉	headings
〈H6〉	headings
〈HEAD〉	head
〈I〉	keywords, parent
〈IMG:SRC〉	images
〈META:CONTENT〉	$NAME
〈STRONG〉	keywords, parent
〈TITLE〉	title
〈TT〉	keywords, parent
〈UL〉	keywords, parent

9. 결론

SOIF는 인터넷 자원을 효과적으로 전송하기 위한 메타데이터로서 파일의 크기, URL 정보, 그리고 대상 객체를 복사할 수 있는 관련된 정보를 제공하며 로봇에 의해 자동 생성하거나 관리자나 저자에 의해 직접 생성이 가능하다는 특징을 가지고 있다. SOIF는 Harvest 시스템에서 사용하기 위해 개발되었지만, 이후 넷스케이프사의 목록 서버 제품이나 스탠포드 대학의 인터넷 검색 프로토콜의 연구 프로젝트인 STARTS(Stanford Protocol Proposal for Internet Retrieval and Search) 등에 응용되었고, 이후 인터넷 자원 전송을 위한 메타데이터 개발에 많은 영향을 끼쳤다.