JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 18, No. 2 : 387~393 Feb. 2014

비정형 대용량 데이터 입력 및 출력 시스템 설계 및 구현

김창수[1] · 심규철[2] · 강병준[2] · 김경환[2] · 정회경[1*]

Design and Implementation of Input and Output System for Unstructured Big Data

Chang-su Kim[1] · Kyu-chul Shim[2] · Byoung-jun Kang[2] · Kyung-hwan Kim[2] · Hoe-kyung Jung[1*]

[1]Department of Computer Engineering, Paichai University, Daejeon 302-735, Korea
[2]Commu Co. Ltd, Daejeon 305-509, Korea

요 약

컴퓨터의 보급에 따라 비정형 대용량 데이터가 범람하고 이를 효율적으로 처리하기 노력이 요구되고 있다. 이에 본 논문에서는 오피스(office) 파일(아래한글, MS-Office 등)에 입력된 데이터를 바로 XML로 변환하고, 사용자가 XML 매핑 파일을 만들어서 워드프로세서에 입력 된 데이터를 바로 추출하여 데이터베이스에 저장하는 시스템을 제안하였다. 또한, 본 시스템은 워드프로세스에 양식을 미리 작성하여 필요한 데이터를 데이터베이스에서 조회하여 워드프로세서 문서를 응용프로그램에서 오피스 파일을 생성 할 수 있다. 이는 대용량의 비정형 데이터를 활용가능하게 할 것이다.

ABSTRACT

In recent years, the spread of computers is increasing, and efficient processing effort for unstructured Big Data is required. In this paper, we are proposed a system to extract the data typed in a word processor quickly by user creating and XML mapping file after converting XML data that has been entered in the office file(HWP, MS-office). In addition, we proposed a system is able to lookup the necessary data from a database by entered form in advance and convert word processor document to office files by the application program. The unstructured big data will be available to be used.

키워드 : 대용량 데이터, 매핑, 비정형 데이터, XML

Key word : Big Data, Mapping, Unstructured Data, XML

접수일자 : 2013. 12. 22 심사완료일자 : 2014. 01. 09 게재확정일자 : 2014. 01. 23

*** Corresponding Author** Hoe-Kyung Jung(E-mail:hkjung@pcu.ac.kr, Tel:+82-42-520-5640)
Department of Computer Engineering, Paichai University, Daejeon 302-735, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2014.18.2.387 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

일반적으로 웹상의 데이터 입력은 정형화된 입력 화면을 제공하고 사용자가 화면에 데이터를 입력하면 적합성 등을 검사하고 이를 데이터베이스에 저장하고, 데이터베이스에 저장된 데이터를 웹 화면에 출력하여 사용자에게 데이터를 제공하고 있다. 그러나 비정형화된 데이터에 대해 웹을 이용하여 데이터를 입력할 수 있는 시스템의 구축이 매우 어렵고 많은 시간이 소요된다.

이에 따라, 일반적으로 사용자가 사용하는 아래한글이나 Microsoft-Office(워드, 파워포인트, 엑셀 등)를 이용하여 다양한 양식으로 데이터를 입력하고, 입력된 파일에 대해 웹을 통하여 업로드 하여, 데이터를 추출하여 데이터베이스에 데이터를 입력할 수 있는 시스템 개발이 요구되고 있다.

이에, 본 논문에서는 업로드 된 바이너리(binary) 파일을 XML(eXtensible Markup Language) 파일로 변환할 수 있는 엔진을 구현하였다. 또한, 변환된 XML 파일과 Mapping XML 파일을 파싱하여 데이터베이스에 데이터를 입력하거나 또는 데이터베이스에 입력된 데이터를 양식(HWP, MS-Office) 파일로 출력할 수 있는 시스템을 설계 및 구현 하였다.

Ⅱ. 관련연구

2.1. 복합 파일(Compound File) 구조

아래한글 및 Microsoft-Office 파일의 구조는 모두 복합 파일 구조를 사용하고 있으며, 내부적으로 스토리지(Storage)와 스트림(Stream)을 구별하기 위한 이름을 가진다. 하나의 스트림에는 일반적으로 바이너리나 레코드 구조로 데이터가 저장되고 스트림에 따라서 압축/암호화가 되기도 한다. 복합 파일을 사용하는 이유는 4기가(giga) 이상의 파일을 저장하기 위하여 사용한다[1].

2.2. HWPML

한글 워드 프로세서 문서를 기술하기 위한 W3C XML 기반의 개방형 마크업 언어이다. HWPML 엘리먼트에 대한 설명은 표 1과 같다[2,3].

처리 과정은 한글 워드프로세서의 원본 문서를 HWPML로 변경 저장하여 매핑(Mapping) XML 파일과 매핑하여 필요한 부분의 데이터를 추출한다.

표 1. HWPML 엘리먼트
Table. 1 HWPML Element

엘리먼트 명				
설명		엘리먼트에 대한 설명		
부모 엘리먼트		부모엘리먼트 명		
자식 엘리먼/ 엘리먼트 값		자식 엘리먼트 값		
속성	속성명1	속성1에 대한 설명	값의 범위	기본값
	속성명2	속성2에 대한 설명	값의 범위	기본값

2.3. 데이터 매핑

매핑은 두 가지의 다른 구조를 지닌 명세 형식에서, 동일하거나 비슷한 의미를 지닌 데이터 요소들을 연결하고, 필요에 따라 연결정의에 추가적인 의미를 부여하여, 두 명세 형식간의 관계를 정의하는 것이다.

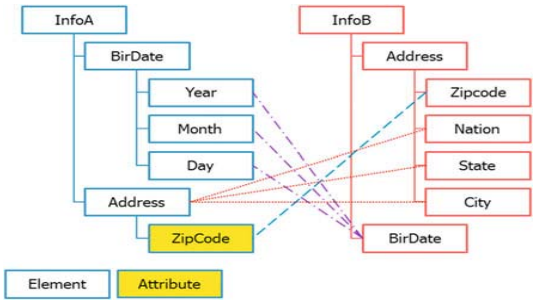


그림 1. 데이터 매핑
Fig. 1 Data Mapping

그림 1은 두 문서 구조 정보에 대해 데이터 매핑을 통해 정의한 예로 써, 보는 바와 같이 두 구조 정보는 같은 내용을 담고 있으면서도 다른 구조를 지니고 있다. BirDate이라는 엘리먼트 하위로 Year, Month, Day를 구분짓는 구조를 n:1 의 관계로 BirDate이라는 하나의 엘리먼트에 매핑하고, Address라는 엘리먼트를 1:n 의 관계로 분리하여 매핑하며, 1:1 의 관계인 속성으로 표현된 Zipcode를 엘리먼트로 표현된 Zipcode와 매핑한 모습이다.

Ⅲ. 비정형 대용량 데이터 입력 및 출력 시스템 설계

3.1. 시스템 구성

비정형 대용량 데이터 입력 및 출력 시스템 구성은 다양한 O/S 환경에서 설치가 가능하도록 개발하고 또한 Active-X 방식을 사용하지 않도록 설계하여 다양한 브라우저를 지원 할 수 있도록 설계하였다. 그림 2는 전체 시스템 구성도이다.

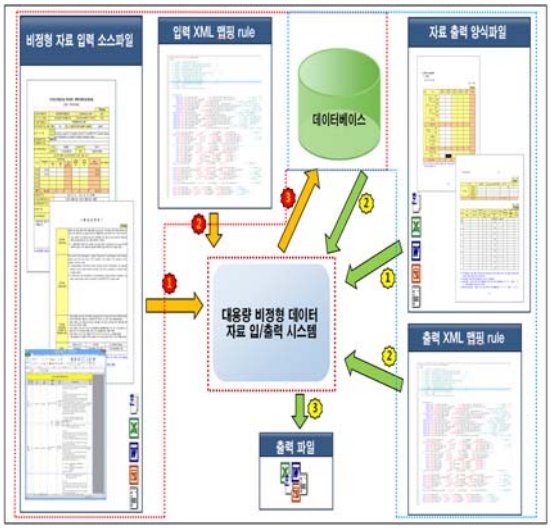


그림 2. 전체 시스템 구성도
Fig. 2 Diagram of System

비정형 대용량 데이터 입력 및 출력 시스템은 비정형 자료와 XML 맵핑 규칙을 입력 받아 문서 구조를 파악하고, 파악된 데이터들 중에 필요한 정보를 추출하여 구조 분석기에서 DOM(Document Object Model) 인터페이스를 이용하여 데이터를 트리 노드에 인덱스 값으로 매핑한다.

본 시스템의 엔진은 크게 2개 부분으로 분리하여 설계하였다. 이는 다른 시스템 구축 시에도 엔진을 각각 활용 할 수 있도록 하였으며, 두 엔진간의 통신은 IPC (Interprocess Communication)를 사용하여 서로 데이터를 송수신 할 수 있도록 하였다.

웹서버를 지원하기 위하여 Open Source 엔진인 NODE JS 엔진을 활용하였다. NODE JS 엔진은 기존 웹서버나 웹 어플리케이션 서버보다 가볍고 JavaScript를 사용하여 개발이 가능하며 다양한 어플리케이션에 활용이 가능하다 최근에는 계속 버전을 업그레이드하여 안전성 및 성능이 많이 향상되었다. 비정형 대용량 데이터 입력 및 출력 시스템 구성도는 그림 3과 같다.

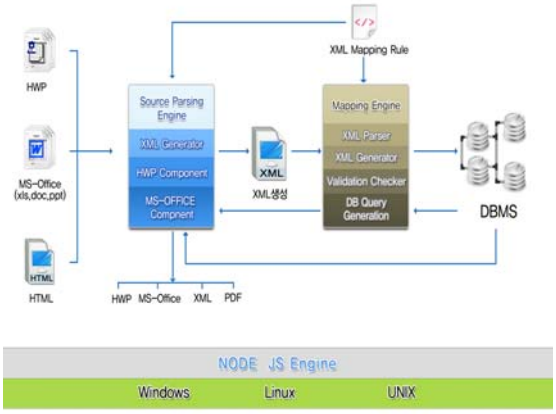


그림 3. 비정형 대용량 데이터 입력 및 출력 시스템 구성도
Fig. 3 Diagram of Input/Output System for Unstructured Big Data

3.2. 원본파일 변환 엔진

원본파일(HWP, MS-Office)을 바이너리 상태에서 바로 XML로 변환 할 수 있다. HWP 파일의 경우 한글과컴퓨터에서 제공하는 HWP 문서 파일 포맷[1]에서 정의된 파일을 원본으로 읽어들여 바로 XML로 변활할 수 있다. 또한, MS-Office 파일의 경우 다양한 오픈 API[4,5]를 이용하여 원본 파일을 바로 XML파일로 변환할 수 있다.

변환단계는 잘 구성된(Well-formed) XML 문서로 변환하는 부분이다. 본 시스템에서 필터링 단계를 거쳐 생성된 중간 단계의 서식 파일을 읽어들인 후, DOM 인터페이스를 이용하여 태그 빌드와 내용 빌드를 생성하고 문서 구조를 검증한 후, 각각의 인덱스 값을 매핑하여 원문서의 변형없이 구조적인 XML 문서를 생성하도록 정의하였다.

3.3. 매핑 XML 파일

본 논문에서는 원본을 XML로 변환하여 이 변환된 XML파일에서 데이터 부분을 추출하기 위하여 매핑 XML 파일을 작성한다. 또한, 원본 XML 파일과 매핑 XML 파일을 파싱하여 원본의 XML 파일에서 데이터

를 추출한다. 추출하기 전에 먼저 원본 XML파일과 매핑 XML 파일의 유효성 검증(Validation check)을 수행하여 XML 파일의 정확성을 확인하여 다음 작업을 진행 하도록 한다.

유효성 검증 후 구조 분석이 이루어지고, DOM 인터페이스를 이용하여 데이터를 필드 단위로 각각 메모리에 적재하여 처리한다. 이는 XML 구문 규칙을 사용하고 있기 때문에 DOM 인터페이스를 제어 할 수 있다.

3.4. 원본파일 파싱엔진 설계

원본파일은 NODE JS 엔진 및 HTTP 프로토콜을 사용하여 서버에 업로드 할 수 있는 기능을 제공하고 HWP 컴포넌트(Component)와 MS-Office 컴포넌트를 사용하여 원본파일을 XML 파일로 변환한다. 또한, 원본파일을 바이너리 파일로 출력에 사용하기 위하여 원본파일을 파싱하여 파일의 변수부분에 데이터를 치환한다. 그리고 원본파일의 복사본을 바이너리 파일 변환하는 기능을 함께 제공할 수 있도록 설계하였다.

변환된 XML파일을 서버의 임시영역에 저장하여 매핑 엔진에 매핑 작업을 의뢰하기 위하여 각종 정보에 대해 통신 모듈을 통하여 전달한다. 원본 파일 파싱 엔진 구성도는 그림 4와 같다.

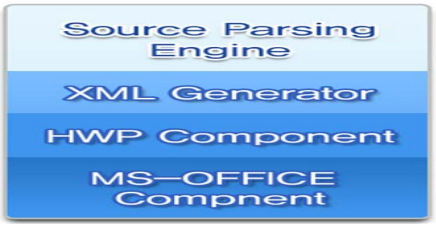


그림 4. 원본 파일 파싱엔진 구성도
Fig. 4 Diagram of Source File Parsing Engine

문서를 입력 후, 원본파일 파싱엔진에서는 문서 전체를 한 라인씩 읽어 들여 각 입력문서의 구성요소를 각각 순차적으로 비교하여 내용이 있으면 그 내용들을 임시 서식 파일에 저장한다.

3.5. 매핑 엔진 설계

매핑 엔진은 XML 파서를 이용하여 원본 XML 파일과 매핑 XML 파일을 검증하여 이상이 없을 때 XML 문서 생성기(XML Generator)를 이용하여 원본파일과 매핑 XML 파일을 분석한다. 다음으로 데이터베이스에 입력할 수 있는 SQL 문장을 생성한다. 만약 검증의 문제나 두 XML 파일에 오류가 발생하면 사용자에게 그 오류 코드를 사용자에게 전송한다. 매핑 엔진은 XML 파싱 및 데이터베이스에 데이터를 입출력할 수 있는 모듈을 모두 포함하여 입력 및 출력에 사용할 수 있도록 독립적인 형태로 설계하였다. XML 매핑 엔진 구성도는 그림 5와 같다.

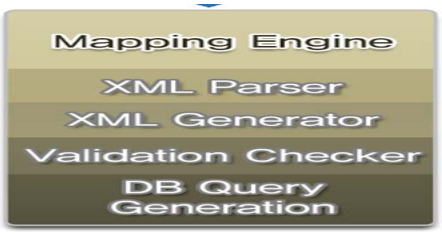


그림 5. XML 매핑 엔진 구성도
Fig. 5 Diagram of XML Mapping Engine

3.6. 문서 생성기

문서 생성기는 구조 분석기 정보를 이용하여 XML 문서를 생성하는 역할을 담당한다. 구조 분석기에서 요구하는 엘리먼트 추출기와 속성 추출기에서는 각각의 엘리먼트와 속성 내용을 추출하여 구조체에 삽입하고, 내부 구조 생성기에서 시작 태그가 생성되기 위한 실제적 내용의 엘리먼트 이름을 엘리먼트 추출기에서 추출하여 생성한다.

생성된 시작 태그 안에 엘리먼트 이름이 같은 속성을 하나씩 추출하여 시작 태그의 엘리먼트 이름 뒤에 속성 내용을 삽입한다. 내용추가에서는 시작태그를 생성 후, 실제적 내용부분을 추가하고, 하나의 엘리먼트가 끝나는 종료 태그를 추가하게 된다. 마지막으로, 예외처리기에서는 엘리먼트나 속성이 표현되지 못했거나, 누락되어진 부분을 표현하도록 하였다.

Ⅳ. 비정형 대용량 데이터 입력 및 출력 시스템 구현

4.1. 시스템 모듈화

본 논문의 시스템은 그림 6과 같이 엔진을 분리하여 개발하였다. 분리된 엔진들은 독립적으로 실행이 가능

하도록 하여 향후 다른 시스템에서 각각의 엔진을 사용할 수 있도록 구현하였다. 또한 엔진에 구현된 컴포넌트들도 독립적인 모듈로 구현하여 다른 어플리케이션에서도 라이브러리 또는 단위 어플리케이션으로 활용이 가능하도록 구현하였다.

4.2. 비정형 대용량 데이터 입력 시스템

비정형 대용량 데이터 입력시스템은 관리자가 원본파일을 배포하여 사용자가 원본파일에 데이터를 작성한다. 시스템에 업로드하면 원본파일과 매핑 XML 파일을 파싱하여 데이터베이스에 데이터를 입력할 수 있도록 구현하였다. 입력된 데이터는 바로 화면에 HTML 형태로 출력하여 필요시 웹 화면에서도 수정이 가능하도록 하였다. 또한, 원본파일을 수정하여 다시 업로드하여 데이터의 내용을 수정 할 수 있도록 구현하였다.

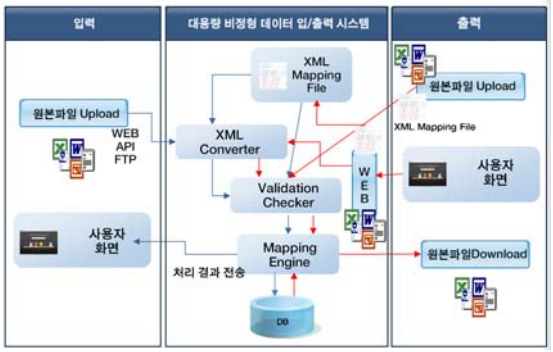


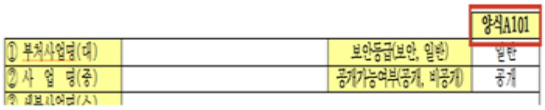
그림 6. 비정형 대용량 데이터 입출력 시스템 구성도
Fig. 6 Diagram of Input/Output System for Unstructured Big Data

4.3. 비정형 대용량 데이터 출력 시스템

대용량 비정형 데이터 출력시스템은 기존 Active-X 방식의 웹 리포팅(Reporting) 툴의 처리방식과는 다르게 처리하였다. HWP, MS-Office의 양식파일을 작성하여 데이터출력 부분에 변수 명을 기입하고 데이터베이스에서 데이터를 조회하여 원본파일의 변수명과 치환하여 다시 원본파일을 작성한다. 작성된 원본파일은 HTTP 프로토콜을 사용하여 사용자에게 다운로드할 수 있는 기능을 구현하여 제공하였다.

4.4. 매핑 XML의 정의

매핑 XML의 내용은 여러 종류의 양식을 구분하기 위하여 formno와 테이블의 데이터를 추출하기 위하여 datapos와 같은 테그를 정의하여 매핑 XML를 정위하였다. 그림 7과 같은 양식을 구분하기 위하여 워드프로세서 파일로 최상위 1칸의 테이블을 만들고 '양식'이라 표기하여 양식을 구분 할 수 있도록 하였다.




```
<양식A101 formno="양식A101" required="true" fmax="1">
```

그림 7. 예제 양식 및 Mapping XML 정의
Fig. 7 Example Form and Mapping XML Definition

원본 파일에서 데이터를 추출하기 위하여 테이블의 위치를 지정할 수 있도록 datapos를 정의하였다. datapos의 위치는 X(수평방향), Y(수직방향)로 지정할 수 있으며 X, Y와 같이 +(칸째), *(건너띄기)의 기호를 함께 사용하여 작성 할 수 있다. datapos 태그작성 유형은 다음 표 2와 같다.

표 2. Datapos 태그 유형
Table. 2 Datapos Tag Type

[유형1] - X+1 : 수평방향 우측으로 1칸 옆의 Cell에서 데이터를 추출함 [유형2] - Y+1 : 수직방향 아래로 1칸째 Cell에서 데이터를 추출함 [유형3] - Y*2 : 데이터가 반복되는 행 같은 경우 해당 헤더의 위치에서 2칸씩 건너뛰어서 데이터를 추출 함 (헤더영역의 Cell이 2줄로 나뉘어 있음) [유형4] - Y+1 : 데이터가 반복되는 행 같은 경우 해당 헤더의 위치에서 1칸씩 건너뛰어서 데이터를 추출 함 (헤더영역의 Cell이 1줄로 되어 있음) (칸수가 1인 경우는 +,* 둘다 같은 결과를 가져옴)

아래 그림 8의 예제는 datapos=Y+1의 매핑 XML정의 예제 부분이다.

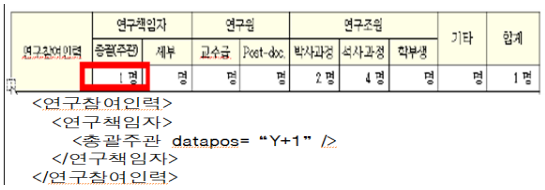


그림 8. 매핑 XML datapos=Y+1 예제
Fig. 8 Example of Mapping XML Datapos=Y+1

Ⅴ. 결 론

본 논문에서 제안한 시스템은 웹상에서 비정형 대용량 데이터에 대해 다양한 양식을 통하여 시스템에 빠른 입출력을 제공할 수 있는 프로그램으로 사용자나 관리자 모두에게 편리한 방식의 데이터 입력 시스템으로 사용할 수 있다. 따라서, 매우 편리하게 사용이 가능 할 것으로 예상된다. 개발자는 다양한 양식을 입력 받기 위하여 프로그램을 개발이나 변경이 거의 없으므로 시스템 구축 시 많은 시간을 단축시킬 수 있을 것이다.

본 논문에서 제안한 시스템은 하나의 독립된 솔루션으로 구축이 가능하여 향후 어플라이언스 서버에 소프트웨어를 탑재하여 단독으로 다양한 포맷의 데이터를 고속으로 처리할 수 있도록 개발이 가능할 것이다.

하지만 이 시스템은 사용자가 매핑 XML 파일을 규칙에 맞도록 작성하여야 하므로 매핑 XML 파일 작성 방법을 습득하여야하는 단점이 있다. 이를 해결하기 위하여 사용자가 매핑 XML 파일을 쉽게 작성할 수 있도록 GUI 어플리케이션을 작성하여 제공한다면 이 단점도 해결이 가능하리라 사료된다.

감사의 글

본 논문은 중소기업청에서 지원하는 2013년도 산학연협력기술개발사업(No. C0115138)에 의하여 이루어진 연구로서, 관계부처에 감사드립니다.

REFERENCES

[1] D Rentz, “Microsoft Compound Document File Format,” [Internet]. Available: http://www.openoffice.org.zaxyproxy.com/.

[2] Hangul and Computer Co., Ltd.. Hangul document file formats Open project [Internet]. Available: http://www.hancom.com/.

[3] J. H. Yun, J. H. Park, and S. J. Lee, “Methods for Investigating of Edit History about MS PowerPoint Files That Using the OOXML Formats,” *Journal of Korea Information Processing Society*, vol. 19, no. 4, pp. 215-224, Apr. 2011.

[4] Apache[Internet]. Available: http://www.apache.org/.

[5] S. M. Han, “Open Source DBMS based Design and Implementation of Query and Transformation Processor for Geo-Spatial Information Metadata,” M. S. dissertation, Hansung University, Seoul, MA, 2010.

김창수(Chang-Su Kim)

1996년 배재대학교 전자계산학과(이학사)
1998년 배재대학교 전자계산학과(이학석사)
2002년 배재대학교 컴퓨터공학과(공학박사)
2005년 ~ 2010년 청운대학교 인터넷학과
2013년 ~ 현재 배재대학교 컴퓨터공학과 조교수
※관심분야 : 멀티미디어문서정보처리, 차세대 인터넷, USN, 모바일 웹서비스

심규철(Kyu-Chul Shim)

2008년 학점은행제 정보통신공학(공학사)
2010년 한밭대학교 컴퓨터공학과(석사 재학)
2010년 ~ 현재 (주)커뮤 S/W 기술연구소 책임연구원
※관심분야 : 멀티미디어 문서정보처리, XML, SVG, HTML5

강병준(Byoung-Jun Kang)

1997년 한밭대학교 전자계산학과(공학사)
2011년 한밭대학교 컴퓨터공학과(공학석사)
2010년 ~ 현재 (주)커뮤 S/W 기술연구소 책임연구원
※관심분야 : 멀티미디어 문서정보처리, XML, SVG,

김경환(Kyung-Hwan Kim)

2000년 배재대학교 컴퓨터공학과(공학사)
2014년 배재대학교 컴퓨터공학과 석사과정
2005년 ~ 2011년 ㈜시스템뱅크 기업부설연구소 연구소장
2011년 ~ 현재 ㈜커뮤 기업부설연구소 연구소장
※관심분야 : 멀티미디어 문서정보처리, XML, SVG, Web Services, Semantic Web

정회경(Hoe-Kyung Jung)

1985년 광운대학교 컴퓨터공학과(공학사)
1987년 광운대학교 컴퓨터공학과(공학석사)
1993년 광운대학교 컴퓨터공학과(공학박사)
1994년 ~ 현재 배재대학교 컴퓨터공학과 교수
※관심분야 : 멀티미디어 문서정보처리, XML, SVG, Web Services, Semantic Web, MPEG-21, Ubiquitous Computing, USN