반도체디스플레이기술학회지 제23권 제2호(2024년 6월)
Journal of the Semiconductor & Display Technology, Vol. 23, No. 2. June 2024.

비정형 데이터를 활용한 지능형 문서 처리 관리에 관한 연구

박경훈*·서광규*†
*†상명대학교 경영공학과

A Study on Intelligent Document Processing Management using Unstructured Data

Kyoung Hoon Park* and Kwang-Kyu Seo*†
*†Management Engineering, Sangmyung Univ., Korea

ABSTRACT

This research focuses on processing unstructured data efficiently, containing various formulas in document processing and management regarding the terms and rules of domestic insurance documents using text mining techniques. Through parsing and compilation technology, document context, content, constants, and variables are automatically separated, and errors are verified in order of the document and logic to improve document accuracy accordingly. Through document debugging technology, errors in the document are identified in real time. Furthermore, it is necessary to predict the changes that intelligent document processing will bring to document management work, in particular, the impact on documents and utilization tasks that are double managed due to various formulas and prepare necessary capabilities in the future.

Key Words : Document management service, Intelligent document platform, Parsing technology, Compilation skills, New document Bert function

1. 서 론

최근 다양한 문서관리 플랫폼의 등장과 Cloud service의 확산으로 창의적 아이디어에 기반한 지능적인 SW 제품 및 서비스를 쉽게 직접 구상하고 개발이 가능한 지능형 SW의 관심이 높아지고 있다[1]. 이에 인간 중심적인 초자동화(Hyper-automation) 분야에서는 단순하고 반복적 비즈니스 프로세스를 담당하는 분야에 관해서 지속적인 연구 및 광범위한 SW가 개발되고 있다. 현재까지의 전자문서 관리의 방식은 단순히 저장하고 불러들여서 직접 업데이트 하는 방식으로 진행되고 있고 담당자 혹은 관리자가 변경된 이후에는 과거 자료와 신규 자료에 대한 유사점 및 차이점을 찾는 데 많은 시간이 소요되고 있다. 또한 문서 작성에 대한 검토를 상위자 또는 자체 필터링을 통하여 문서를 검증하는 데 많은 시간이 소요되고 있고 여러 수식이 들어가는 문서일 경우에는 한글파일에서 작업한 내용을 다시 엑셀이나 다른 수식 프로그램을 통하여 계산식을 새롭게 수립하거나 별도로 관리해야 하는 어려움이 있다. 이번 연구에서는 보험약관, 산출방법서, 사업방법서 등과 공공기관의 법, 규정, 계약 등의 법령 개정 시 필요한 개정문 등 여러 전자 문서를 효율적으로 관리하기 위한 방법을 제시하고 이를 보험약관 문서의 적용하여 지능형 문서처리 기술에 대한 객관성을 확보하였다.

†E-mail: kwangkyu@smu.ac.kr

2. 연구 배경

2.1 선행연구 조사

본 연구 주제인 문서처리 시스템 관련 선행연구를 살펴보면 다음과 같다.

먼저 “FPGA를 이용한 하드웨어 기반 고성능 XML 파

싱 기법”에서는 XML(eXtensible Markup Language) 파싱 성능을 높이기 위해서 FPGA를 이용하여 파서를 설계하여 검증하였다. 이는 구조화 된 표준문서에서 실시간 XML 파싱 방안을 제시하였다[2].

“문서 유사도를 통한 관련 문서 분류 시스템 연구”에서는 머신 러닝 기술을 이용하여 문서를 분석하고 이를 바탕으로 문서를 분류하는 방법을 제시하였다[3].

“텍스트 분석 기술 및 활용 동향”에서는 다양한 유형의 데이터 중 특히 텍스트 분석에 대한 연구 사례를 소개하였고 텍스트 분석 수행 방법에 일반적인 방법과 텍스트 분석 기술을 제시하였다[4].

“문서구조 추출기법을 이용한 엔지니어링 문서 텍스트 정보의 XML 변환”에서는 계층구조가 복잡한 형식을 띄는 엔지니어링 문서의 비구조화된 텍스트 정보를 계층구조에 따른 준 구조화된 XML문서로 변환하여 문서구조 분석 기법을 보여주는데 이는 텍스트 정보 분석을 통해서 구조화시키는 작업이 중요한 부분을 차지한다[5].

“문서의 계층화를 이용한 문서비교 방법”에서는 텍스트 문서의 효율적인 검색 방법 중 문서와 비슷한 문서를 의미적으로 찾아내기 위한 계층화 방법을 제시하였고 이는 개념 일치도를 이용하는데 많이 사용되고 있는 방법이고 본 연구도 계층화를 통한 방법으로 문서 비교를 진행하였다[6].

“심층신경망을 이용한 PCB 부품의 인쇄문자 인식” 에서는 인쇄 문서에서의 텍스트 추출 및 문자를 인식하는데 딥러닝으로 문자 분리 없이 문자를 연속적으로 이용하는 LPRNet(License Plate Recognition via Deep Neural Networks) 통해서 인쇄 문자의 고속, 고정도 인식을 진행하였다[7].

기존 연구에서는 HWP 및 다국어를 지원하는데 한계가 있었으나, 본 논문에서는 한글 띄어쓰기, 단어교정, 문구교정, 문장 일괄변경, 신구대조, 개정문서 및 문서 버전관리가 가능하도록 하고 문서의 비정형데이터에 대한 자동 수식 계산을 통해서 보험료에 대한 책임준비금(Policy reserve) 산출, 문서 작성 및 비교의 자동화, 약관, 상품설명서의 문서적 오류 점검 정확도를 한층 강화한 측면에서 기존의 선행연구와 차별성을 갖는다.

2.2 약관문서의 표준 인덱스 정의

비정형 데이터 추출을 하기 위한 초기 데이터의 확보를 위해서 보험사 공시실을 통해 제공하고 있는 약관 PDF 문서를 수집하기위해 협회 및 각 보험사의 공시실에서 제공하는 판매중인 혹은 판매 중지된 상품의 약관 문서를 활용하였다. 문서의 종류는 크게 3가지를 기본으로 파싱(Parsing) 하였으며, 약관, 산출방법서를 근거로 문서 레파지토리와 문서의 구조체를 구조화 하였다. 국내의 모든 정형화 되어있는 모든 문서는 편-장-절-관-조-항-호-목으로 구분이 된다. 이를 표준화 인식하여 문서를 상기의 구분으로 구성하고 있으나 예외가 되는 많은 문서가 존재하고 있으며, 일반 기업의 문서는 문서의 가장 기본이 되는 표준 인덱스 구분 조차 많은 차이를 갖고 있다. 이러한 과정에서 규정, 약관, 법률에 한해서는 많은 부분이 표준화가 이루어져 있으며, 특히 약관의 경우는 사용자에게 전달하는 최종 계약 후 문서로써 의미가 있으며, 이를 본 연구과제의 핵심 문서로써 선정하여 진행하였고 그 외 기타 문서는 예외 문서로써 파싱과 수집, 인덱싱 등에서 약간의 차이와 예외를 구분하여 진행하였다. 수집한 약관 문서의 구조를 파악하기위해 문서 구조의 표준 인덱스 공통 규칙을 찾아내는 과정을 거치고 표준 인덱스 정의는 약관 문서를 수집한 후 각 약관을 구성하는 문서 구조를 파악하고, 표준규격을 찾아내 보았다. 표준규격으로 정의된 인덱스 정규화와 인덱스 별 문서 내에서 선언된 설정값을 정의하여 정규화된 표준 인덱스를 찾아낼 수 있다.

3. 문서 파싱 방법

3.1 약관문서 파싱

보험약관에서 추출된 본문내용을 Fig. 1 절차를 통해 문장 단위로 인덱스 및 문서 구조를 분석한다. 문장 별 파싱은 문장 별로 미리 선언한 표준 인덱스에 적합한지 판단하여, 해당 문장의 인덱스를 결정하는 과정이다. 이렇게 결정된 인덱스는 앞/뒤 문장의 인덱스와 결합하여 상/하위 구조를 판단할 수 있다. 상/하 구조를 판단한 각 문장은 트리구조 형태로 문서의 구조를 구성할 수 있다. 트리형태로 구성된 문서 구조를 통해 누락 혹은 중복된 인덱스를 판단하고, 이를 오류로 출력할 수 있다. 이렇게 구

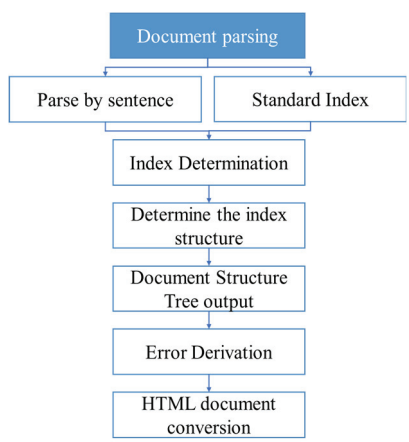

Fig. 1. Document structure analysis procedure.

조화된 본문내용을 각 인덱스 별 스타일에 맞게 CSS (Cascading Style Sheet) 및 태그를 적용하여 HTML(Hyper Text Markup Language) 문서로 변환 시키면 활용 가능한 데이터 확보가 가능하다.

3.2 문서 개정

전자 문서의 변경대비표는 문서 개정 시 이전 버전의 정보와 현재 수정된 버전의 변경된 부분에 대해 한눈에 볼 수 있도록 해야 사용자가 편리하게 사용할 수 있기 때문에 표 형태가 기본으로 제공되도록 하였다. 또한 HWP, DOC, PDF, TXT 문서의 구조와 관계없이 문서의 웹에서 비교 결과를 확인 할 수 있도록 실시간으로 변경대비표 생성, 실시간 Database 연동, 표준 문서 템플릿 제공, 자동 Wiki 생성하여 사용자 요구사항에 따라 만들 수 있도록 구현하여 As-Is와 To-Be의 내용을 문서상에 색상으로 표현하여 구분해 보았다. 이를 위해서 파싱 후 핵심 키워드의 인덱싱 구성을 통하여 빠른 검색이 가능하며, 향후 이미지, 동영상, 음성 등의 검색 기능의 제공을 위한 아키텍처 구성 적용을 통해서 웹 페이지로 제공을 하였다. 웹 에디터는 HTML문서 형태로 관리되어 있으며, 내용에는 다양한 기능 및 스타일을 제공하기 위해 복잡한 구조로 구성되어 있다. 그러나 변경대비표는 문서의 기능이나 스타일은 제외하고, 문서의 구조와 텍스트, 이미지의 개정내용만을 표기 하기 때문에 문서의 단순화하는 과정이 필요하다.

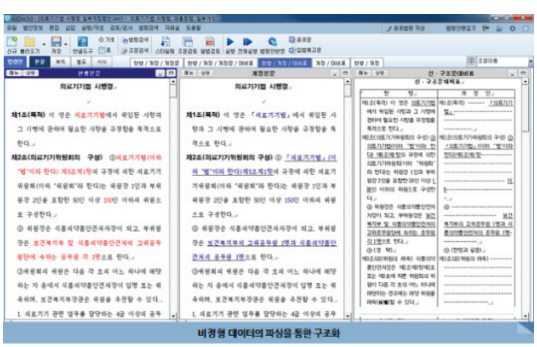

Fig. 2. Document comparison screen through parsing.

이를 통해서 변경 이전 문서와 변경 대상 문서를 구조화하여 트리 형태로 구성한 후 각각의 매칭되는 노드를 찾아 연결한다. 연결방법에는 우선 하위까지 동일한 문장 > 제목은 동일하고 하위는 2/3이상 동일한 문장 > 하위 제외하고 해당 문장만 동일한 문장 > 연결되지 않고, 앞/뒤 문장이 연결되어 연결 안 된 문장의 수가 동일한 문장의 순으로 연결을 진행한다. 문장 간 연결 후 내용이 일치하지 않은 문장을 추려 개정 범위를 분석하고 개정 범위 분석은 개정 범위 분석을 위해서는 조건이 필요한데, 아래 Table 1 기준에 따라 적용을 진행하였다.

Table 1. Revision scope analysis conditions

Revision Scope	Explanation	Conditions
Partial Revision	Revise portion of a sentence	
Full Revision	Revise the entire sentence	- Number of revised words segment: 2/3 or more - Revise one or more of two or more-word segment - Add unnumbered context at the bottom
Delete	Delete existing sentence	
Create	Add new sentence	
Shift	Moving numbers and in between numbers	- Within the same type of number - Context without a number will not be moved

3.3 산식 추출 및 자동계산 방법

문서의 문맥, 내용, 상수, 변수 등을 자동분리하기 위해서는 Lexical 분석, Syntax 분석, Intermediate code 생성, Code Optimizer, Target code Generator를 통해서 모든 데이터를 수집한다.

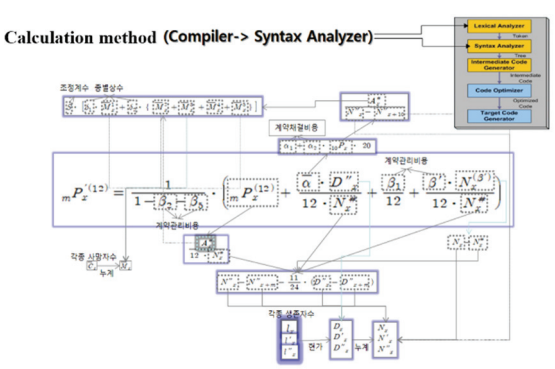

Fig. 3. Parsing and compilation techniques.

Fig. 3처럼 각각의 데이터는 보험약관의 수익성 분석에 관한 사항을 통해서 그 값을 지정하여 준다. 산출방법서에는 「보험업법」 제120조제1항에따라 보험 종목 별 또는 계약기간 경과 별로 책임준비금 계산한 보험료 적립금과 미경과 보험료를 정하고 있습니다. 이와 같은 값을

확인하기 위해서는 Fig 4 에 제시한 산출방법서에 기술된 산식을 분석하여 해당 보험종목의 보험료를 산출할 수 있다. 보험사에서는 보험종목마다 복잡한 산식을 계산하기 위해 별도의 산출 엑셀의 형태로 엑셀 문서를 관리한다. 이 산출 엑셀 문서는 종목별로 산출할 수 있는 산출식을 스크립트로 계산해 놓은 형태이다. 본 연구에서는 이를 산출 엑셀이 아닌 문서에서 직접 산출식을 도출하고 계산식을 입력한 정보와 매핑하여 자동으로 계산하도록 개선하였다. 위와 같은 내용은 보험약관의 수익성 분석에 관한 사항을 참고하면 그 값을 지정하는 것은 어렵지 않게 적용 할 수 있다. 이러한 서식을 계산을 위해서는 보험종목별로 가입 나이, 성별, 보험기간, 납입기간 등의 범위를 나타내는 데이터 인 모델포인트도 함께 추출해야 한다. 모델포인트는 가입 가능한 연령과 성별을 기준으로 계약한 보험기간 및 납입 기간에 따라 결정되는 보험료를 산정하기 위해 정의된다. 고객이 보험에 가입시 본인에 맞는 모델포인트를 찾아 미리 계산된 보험료로 계약을 진행하는데 활용한다.

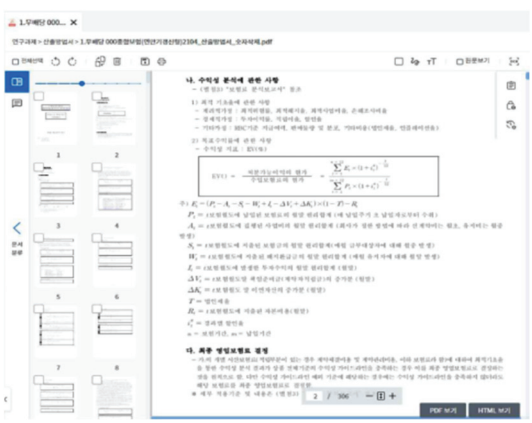

Fig. 4. Insurance premium calculation method.

보험료 계산을 위해서는 마지막으로 위험률 데이터를 찾아야 한다. 보험료를 산출하는데 사용되는 위험률은 실제 발생률이 항상 일정하지 않고 변동 할 수 있기 때문에 실제 발생률에 통계적 안전 할증을 더하여 산출된다. 최종 자동 계산을 위한 이자율과 가입자수, 기준가입금액 및 모델포인트범위를 입력을 마친 후면 로직을 통해서 자동으로 Fig5와 같이 보험료 계산을 진행할 수 있다. 보험료 계산은 이자율과 가입자수를 대입하여 각 가입 나이 및 성별, 보험기간, 납입 기간에 따라 보험료가 결정되고 결정된 보험료에 기준가입금액을 곱하여 월별로 실납부할 보험료가 계산된다.

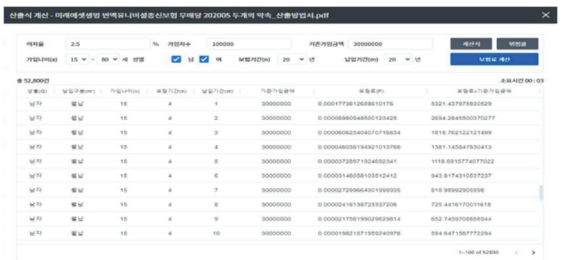

Fig. 5. SW with final insurance premium calculated.

4. 결 론

본 연구에서는 파싱 기술의 고도화 기술인 컴파일 기술을 활용하여 전자문서 비정형 데이터에 수집과 관리가 가능하도록 하여 이러한 데이터의 자동 수집을 통해서 문서를 자동으로 업데이트하고 문서의 정보를 정확하게 읽음으로 인하여 또 다른 고급 문서의 정보를 받을 수 있도록 하였다. 이를 통해 다음과 같은 결론을 도출할 수 있다.

1) 문서의 자동으로 해석 및 가능한 분야로 서비스 확장 가능
2) 초기 문서의 정보 제공하면 자동 업그레이드를 통한 지능형 문서 제공
3) 문서 관리 작업 시간의 단축 및 업무 향상 결과로 비용 절감 및 업무 효율화
4) 인공지능의 발전을 통하여 은행, 보험 업종의 문서, 문자 자동화 서비스 업 증가로 AI 반도체 증가 및 관련사업 성장
5) 개인 정보를 오픈을 통해 또 다른 고급 문서의 정보를 받을 수 있는 오픈 플랫폼으로 확장이 가능
6) 개인 문서의 히스토리 정보를 맞춤형으로 관리 가능함으로써 문서의 통함 관리 구현이 가능해지고 수식이나 비정형 데이터가 있는 문서가 통합 관리

향후 연구로는 본 연구에서 제시한 방법을 활용하여 개인 문서 관리 적용을 위해서는 사용자가 활용 가능한 개인별 클라우드 서비스 제공을 위한 모델 개발과 이를 활용한 초기 모델의 AI 서비스 확장에 대한 서비스 품질 테스트가 필요하다.

감사의 글

본 논문은 2024년 상명대학교 교내연구비를 지원받아 수행하였음.

참고문헌

1. H. C. Jung, K.-K. Seo, “Data Standardization Method for Quality Management of Cloud Computing Services using Artificial Intelligence”, Journal of the Semiconductor & Display Technology, Vol. 21, No. 2, pp. 133-137, 2022.
2. Kyu-hee Lee, Byeong-seok Seo, “Hardware-Based High Performance XML Parsing Technique Using an FPGA”, The Journal of Korean Institute of Communications and Information Sciences, Vol.40 No.12, pp.2469-2475, 2015.
3. Jisoo Jeong, Minkyu Jee , Myunghyun Go, Hakdong Kim, Heonyeong Lim, Yurim Lee, Wonil Kim, “Related Documents Classification System by Similarity between Documents”, Journal of broadcast engineering, Vol.24, No.1, pp.77-86, 2019.
4. Namgyu Kim, Donghoon Lee, Hochang Choi, William Xiu Shun Wong. “Investigations on Techniques and Applications of Text Analytics”, The Journal of Korean Institute of Communications and Information Sciences, Vol.42, No.02, pp.471-492, 2017.
5. Sangho Lee, Junwon Park, Sangil Park, Bonggeun Kim, “Transformation of Text Contents of Engineering Documents into an XML Document by using a Technique of Document Structure Extraction”, Journal of the Korean Society of Civil Engineers, Vol.31,No.6, pp. 849-856, 2011.
6. Myunggwon Hwang, Hyunjang Kong, Kwangsu Hwang, Pankoo Kim, “The Method of Document Comparison using Document Hierarchy”, Communications of the Korean Institute of Information Scientists and Engineers, pp.143-147, 2006.
7. Tai-hoon Cho, “Recognition of Characters Printed on PCB Components Using Deep Neural Networks”, Journal of the Semiconductor & Display Technology, Vol. 20, No. 3. pp. 6-10, 2021.

접수일: 2024년 5월 28일, 심사일: 2024년 6월 17일, 게재확정일: 2024년 6월 21일