

모바일 교환권 통합 관리를 위한 확률적 예외기반 텍스트 제거 알고리즘의 설계 및 구현

김찬솔*·정서현*·김다영**·박시은**·우예은**·이강민**·지윤서**·이채원**·함현지**·차주호***

Design and Implementation of Probabilistic Exception-Based Text Elimination Algorithm for Integrated Mobile Voucher Management

Kim,Chansol·Jeong,Seohyeon·Kim,Dayoung·Park,Sieun·Woo,Yeeun·Lee,Kangmin·Ji,Yoonseo·Lee,Chaewon·Ham,Hyeonji

〈Abstract〉

This paper presents a probabilistic exception-based text removal algorithm designed for mobile gift voucher management systems. This algorithm optimizes data extraction from mobile voucher images using Clova OCR and regular expressions implemented in a PHP server environment. In addition, we apply a unique weighting formula to separate valuable data and discard irrelevant information to increase system efficiency. The successful results of web application tests contribute to this field by not only ensuring the effectiveness of the algorithm, but also demonstrating its potential applicability in a variety of text-based systems.

Key Words : Exception-Based Algorithm, Optimized Data Extraction, Image Data Extraction, Probabilistic Algorithm, Voucher Management

I. 서론

현대 사회는 디지털화의 파도 속에서 스마트페이(Smart Pay)와 간편결제 서비스의 이용이 일상화되는 시대를 맞이하고 있다. 특히 스마트폰을 활용한 금융 거래

는 현금과 실물 카드를 대체하는 추세로, 이러한 변화 속에서 모바일 교환권의 활용도가 눈에 띄게 증가하고 있다. 바코드 인식을 통해 특정 상품 구매가 가능한 모바일 교환권은 선물이나 금액 할인 등 다양한 목적으로 활용되고 있으며, 그 거래량은 날이 갈수록 증가하는 추세이다. 현재 모바일 교환권 관리 서비스는 각 발급처의 전용 애플리케이션에서 제공되고 있으며, 일부 모바일

* 청운대학교 공과대학 멀티미디어학과 학부생(제1저자)

** 청운대학교 공과대학 멀티미디어학과 학부생(공동저자)

*** 청운대학교 공과대학 멀티미디어학과 교수(교신저자)

지갑 서비스를 제공하는 애플리케이션은 모바일 교환권의 이미지를 업로드하여 다양한 발급처의 모바일 교환권을 한 곳에서 모아 관리하는 서비스를 제공하고 있다.

이러한 서비스는 모바일 교환권을 한 곳에 모아 두어 쉽게 볼 수 있게 하고, 발급처에 구애받지 않고 사용자 인증을 통해 모바일 교환권을 등록할 수 있게 하는 등의 이점을 제공한다. 또한, 모바일 교환권을 통합하여 관리하고, 종류별, 기간순 등으로 정렬하며, 사용기간에 따른 선택적 알림 기능을 제공함으로써 사용자의 편의성을 크게 향상시키고 있다.

그러나 이러한 편의성을 더욱 높이기 위해서는 모바일 교환권 이미지에서 특정 문자를 인식하여 분리한 후 필요에 따라 사용할 수 있도록 추출하는 알고리즘의 개발 선행이 요구된다. 이런 요구를 충족시키기 위해, 본 논문에서는 딥러닝(OCR)과 특정 문자열 추출을 활용하여 모바일 교환권에서 필요한 정보를 추출하고, 이를 데이터베이스에 동적으로 기록하며, 사용자가 등록하는 교환권에 따라 자동으로 업데이트하는 알고리즘을 연구하고 제안한다. 이를 통해, 본 논문은 기존의 모바일 교환권 관리 방식의 한계를 극복하고, 사용자의 편의성을 더욱 향상시키는 방안을 제시하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 광학 문자 인식(OCR, Optical Character Recognition)을 활용하여 특정 단어 및 문장 추출에 관한 이전 연구를 검토하고, 3장에서는 모바일 통합 관리 및 맞춤 알림 시스템을 위한 확률적 예외기반 텍스트 제거 알고리즘에 대하여 기술한다. 4장에서는 3장에서 설계한 확률적 예외기반 텍스트 제거 알고리즘의 구현에 대해 기술하고 테스트를 수행한다. 5장에서는 결론 및 향후 개선 사항에 대해 기술한다.

II. 관련 연구

광학 문자 인식(OCR)[1] 기술은 다양한 분야에서 활

용되며 그 중요성이 점차 증가하고 있다. OCR은 이미지, 스캔된 문서, 손글씨 등에서 텍스트를 인식하고 추출하는 기술로, 이를 통해 디지털 텍스트로 변환할 수 있다. 이렇게 변환된 텍스트는 컴퓨터가 처리하거나 편집할 수 있게 된다.

OCR 기술의 활용 분야는 광범위하다. 문서 자동화 분야에서는 OCR을 통해 종이 문서를 디지털화하고, 이를 편집 가능한 형태로 변환하는 프로세스가 이루어진다[1]. 이를 통해 문서 관리와 검색이 용이해지며, 종이 문서의 공간 점유를 줄일 수 있다. 또한, 도서관이나 연구 기관에서는 OCR을 활용하여 책이나 문서를 디지털 형태로 변환하고 있다. 이를 통해 오래된 자료를 보존하고, 대중에게 공유하며, 효율적으로 검색하는 것이 가능해진다. 또한, OCR은 번역 분야에서도 활용되고 있다. 외국어 텍스트가 포함된 이미지를 스캔하고, 이를 디지털 텍스트로 변환하여 번역하는 작업에도 OCR이 사용된다[2]. 시각 장애인이나 학습 장애인을 위해, OCR은 인쇄된 문서나 전자 도서를 음성으로 변환하는 데 활용되고 있다. 이외에도 교통 관리와 법 집행 분야에서는 OCR을 이용하여 차량의 번호판을 인식하는 작업이 이루어지고 있다[3].

따라서, OCR 기술은 그 활용 영역이 매우 다양하며, 이는 미래에도 계속해서 확대될 것으로 예상된다. 이러한 다양한 활용성 덕분에 OCR에 대한 연구와 개발은 매우 활발히 진행되고 있으며, 이는 다양한 산업 분야에서의 효율성과 생산성 향상을 촉진하고 있다.

최근의 OCR 기술은 다양한 변수에 대한 고려를 통해 뛰어난 문자 인식 정확도를 보여주고 있다. 이는 Naver, Google, Amazon 등의 선도적인 기업들이 지속적으로 딥러닝 학습을 통해 업데이트를 진행하고 있어 OCR 기술의 정확도는 더욱 보장받을 수 있다. OCR 기술은 다양한 형태로 존재하며[4], 그 중 템플릿 OCR은 이미지 내 특정 서식이나 구역을 명시적으로 지정하여 해당 영역의 텍스트만을 추출하는 방식을 사용한다. 이러한 접근법은 이미지 내 정보의 위치가 일정한 경우 유용하다. 그러나

모바일 교환권과 같은 경우에는 발급처나 발급 형식에 따라 텍스트의 위치가 변동되거나, 발급처의 모바일 교환권 형식이 업데이트 되면서 정보의 위치가 변경되는 경우가 많다. 따라서 일반적인 이미지에서 모든 문자를 인식하는 표준 OCR 기법을 활용해야 한다.

본 연구에서는 모바일 교환권 통합 관리 시스템을 구축하기 위해 Naver의 Clova OCR을 사용한다. Clova OCR은 Naver에서 개발한 광학 문자 인식 기술로, 다양한 폰트와 배경에 대한 높은 인식률을 보여주는 것이 특징이다. 이 기술은 딥러닝 알고리즘을 통해 계속해서 업데이트되고 있으며, 이를 통해 향상된 인식 정확도를 제공한다. 또한 2018년에 국제 패턴 인식 협회(IAPR)가 주관하는 ICDAR 2013 challenge에서 94.02%의 인식률로 1위를 차지하였다[5].

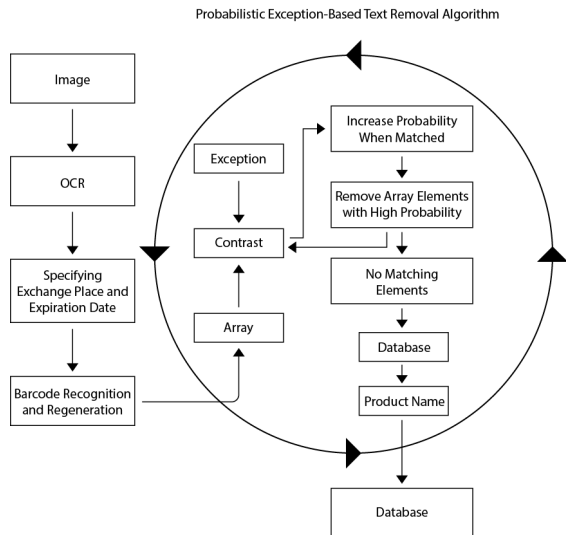
III. 확률적 예외기반 텍스트 제거 알고리즘

3.1 알고리즘 설계

본 연구는 모바일 교환권 통합 관리 시스템에 필요한 데이터화 과정의 효율성을 극대화하는 것을 목표로 한다. 이를 위해 모바일 교환권의 이미지 정보 중 유효기간, 교환처, 상품명 등 시스템에 필수적인 데이터를 효율적으로 변환하고자 한다. Clova OCR을 이용하여 이미지로부터 문자 정보를 추출하고, 불필요한 문자 정보는 예외 처리를 통해 제거함으로써 시스템에 필수적인 값만 세밀하게 확보한다. 이를 통해 사용자의 직접적인 데이터 입력을 최소화하고, 시스템의 동적인 작동을 가능케 하는 것이 본 확률적 예외기반 텍스트 제거 알고리즘의 목표이다.

예외 처리 기준은 다음과 같다. 먼저 이미지에서 인식된 바코드를 분석하여 바코드 번호와 바코드 포맷을 추출하고 재생성한다. 그 다음 Clova OCR을 활용하여 이

미지로부터 얻은 문자 정보에서 정규식을 이용하여 유효기간을 찾아낸다. 이후 카카오톡 선물하기에서 판매 중인 모바일 상품권의 교환처 정보를 활용하는데, 이는 선물하기 시장에서 약 70%의 점유율을 가지고 있음을 고려하였다[6]. 2023년 11월 9일 기준으로 카카오톡 선물하기에서 판매 중인 모바일 상품권의 교환처는 총 625곳이며, 이들의 상호명 데이터를 이용하여 교환처를 정확하게 식별한다. 마지막으로 상품명과 혼동될 수 있는 단어를 제외한 나머지 정보와 불필요한 특수 기호들을 예외 처리 판단의 기준으로 삼는다.



<그림 1> 확률적 예외기반 텍스트 제거 알고리즘 프로세스 다이어그램

<그림 1>은 모바일 교환권 통합 관리 시스템에 적용된 확률적 예외기반 텍스트 제거 알고리즘의 처리 과정을 시각적으로 표현한 것이다. 이 알고리즘이 포함하는 주요 절차는 다음과 같다.

1. 사전 정보 추출 단계: 본 과정에서는 필수적인 데이터 중 상품명을 제외한 교환처, 유효기간, 바코드번호에 대한 데이터를 사전에 추출한다. 교환처는 모바일 교환권의 교환처 데이터를 활용하여 식별한다. 유효기간은 정규식을 활용하여 다양한 형태의 날짜 표현 문자열들을 찾아내고, 이를 Date 객체로 변환

- 한다. 이 중에서 교환권 구매일과 유효기간이 함께 포함되어 있는 경우를 고려하여, 가장 이후의 날짜를 찾아 반환한다. 바코드번호는 Quagga.js API[7]를 통해 인식된 바코드 번호와 바코드 형식을 jsbarcode API[8]를 통해 재생성하여 데이터화한다.
2. 대조 단계: 주된 목표는 대조군과 실험군의 정보를 분석하여 확률적으로 일치하는 요소를 식별하는 것이다. 이 단계에서는 각 실험군 문자열이 대조군 문자열과 비교되며, 일치하는 문자열이 발견될 때마다 해당 문자열에 점수를 부여한다. 점수는 일치하는 문자열의 순서에 따라 제공하여 더해지며, 이는 문자열의 순서가 예외 처리 판단에 중요한 역할을 하게 한다. 이 과정을 통해 가장 높은 점수를 가진 문자열이 확률적 선택자로 선정된다.
 3. 예외기반 확률적 판단: 대조 단계를 거친 후, 확률적 선택자를 이용하여 원본 배열에서 예외적인 문자열들을 제거하는 작업을 수행한다. 이 과정에서 확률적 선택자가 특정 임계값을 초과하면, 해당 문자열은 원본 배열에서 제거된다. 이 과정은 재귀적으로 반복되어 완전히 제거될 때까지 수행된다.
 4. 상품테이블 등록: 모든 과정이 완료되면 최종적으로 추출된 상품명의 유효성을 검사한다. 상품명이 이미 상품테이블에 등록되어 있는 경우에는 사용자에게 보다 정확한 정보를 제공한다. 반면에 상품명이 상품테이블에 처음 등록되는 경우에는 새로운 상품명을 상품테이블에 등록하여 데이터를 축적한다. 이를 통해 지속적으로 상품명 추출의 정확도가 향상된다.

<표 1>은 확률적 예외기반 텍스트 제거 알고리즘의 입출력, 가중치 설정, 세부 사항을 도표로 기술한 것이다. 알고리즘의 기본 가중치는 (2^{weight}) 로 설정하였다. 다음절에서 기술하는 수식(1)의 연속 일치 가중치 계산 공식(Consecutive Match Weight Calculation Formula)과 같이 consecutiveMatchCount는 연속적으로 일치하는 문자의 개수를 나타내어 연속 일치시 가중치를

지수적으로 증가시킨다.

<표 1> 확률적 예외기반 텍스트 제거 알고리즘

Input	
Control group	A set of strings or words of a specific length n
Experimental group	A set of strings or words of a specific length m that need to be exception-handled
Output	
Removed Strings	A set of strings or words that have been exception-handled and removed from the control group
Remaining Strings	The remaining set of strings or words in the control group after exception handling has been applied
Weight Setting	
Assigning weights every time each string of the control group and the experimental group matches Weight calculation formula: Weighted Count = $\sum (2^{\text{weight}})^{(\text{consecutiveMatchCount})}$	
Detail	
대조군과 실험군의 문자열을 비교하며, 일치하는 문자열이 발견될 때마다 해당 문자열에 가중치를 부여한다. 이때, 일치하는 문자열이 연속적일수록 가중치는 Consecutive Match Weight Calculation Formula 공식에 따라 증가한다. 가장 높은 점수를 가진 문자열이 확률적 선택자로 선정되고, 일치하지 않는 문자열은 제외한다. 다양한 경우를 고려하여 알고리즘이 구조적 유사성을 기반으로 예외 처리를 수행하며, 이를 통해 보다 정확한 정보를 출력한다.	

3.2 알고리즘의 가중치 설정 방법

확률적 예외기반 텍스트 제거 알고리즘은 복잡한 텍스트 데이터 분석 작업의 수행 과정에서 텍스트의 구조적 유사성을 핵심 요소로 취급한다. 이 알고리즘은 일치

하는 문자열의 연속적인 개수를 기반으로 가중치를 계산하며, 이는 텍스트의 일치성을 보다 정밀하게 측정하는데 결정적인 역할을 한다. 본 알고리즘에서 사용하는 가중치의 계산 방식은 수식(1)과 같다.

$$\text{Weighted Count} = \sum (2^{\text{weight}})^{\text{consecutiveMatchCount}} \quad (1)$$

여기에서 \sum 는 각 일치하는 문자에 대한 합계를 의미하고, consecutiveMatchCount는 연속적으로 일치하는 문자의 개수를 나타낸다.

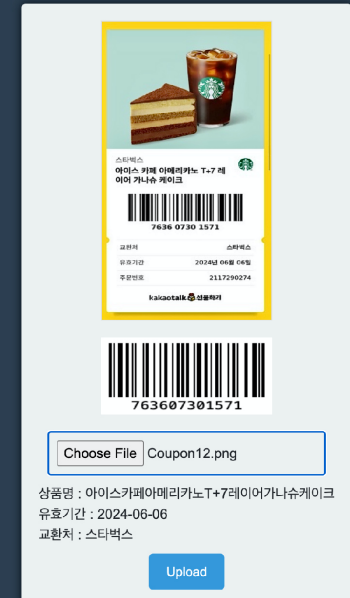
확률적 예외기반 텍스트 제거 알고리즘은 단순히 문자열이 다른 문자열과 얼마나 비슷한지를 정량적으로 측정하는 것을 넘어, 문자열 내부의 문자들이 얼마나 연속적으로 일치하는지를 중요하게 본다. 이를 통해 문자의 순서와 구조가 유사한 텍스트를 더 높은 점수로 평가하게 되며, 이는 결국 예외 처리 판단에 중요한 역할을 한다. 따라서 본 알고리즘의 가중치 부여 방식은 텍스트의 구조적 유사성을 고려하며, 이를 통해 텍스트의 유사성을 보다 정밀하게 측정하는 데 기여한다. 이러한 방식은

텍스트 기반의 예외 처리나 정보 추출 작업에서 높은 효율성과 정확도를 보장한다.

IV. 구현 및 테스트

3장에서 모바일 교환권 통합 관리 시스템에 적용한 확률적 예외기반 텍스트 제거 알고리즘을 기술하였다. 이 알고리즘은 PHP 환경의 서버에서 구축된 테스트 페이지를 통해 클라이언트와 서버를 효율적으로 분리하였다. 서버에서는 데이터 처리를 담당하며, AJAX 통신을 통해 처리 결과를 클라이언트로 전달하였다. 이러한 구현 접근 방식은 모바일 교환권 통합 관리 시스템의 구축 과정에서 중요한 역할을 수행하며, PHP 환경 서버에서의 테스트 페이지를 통한 클라이언트와 서버의 분리, AJAX 통신을 활용한 데이터 전달 등의 기술적인 요소는 시스템의 성능 평가와 검증을 보다 정확하게 수행할 수 있도록 도움을 준다.

<그림 2>와 <그림 3>은 확률적 예외기반 텍스트 제거



Test Log

NAME	Log
11 release	
SortedArray_Probabilistic	교환처 스타벅스
	스타벅스
	아이스 카페 아메리카노 T+7 레
	레이어 가나슈 케이크
	7536 0730 1571
Probabilistic	kakaotalk 선물하기
	84
	0
	0
	0
	4
	0
	1

Probabilistic Selector: -> null
해당 문자열은 존재하지 않습니다
최종배열: 스타벅스, 아이스 카페 아메리카노 T+7 레, 이어 가나슈 케이크
Exchange_Location: 스타벅스
유효기간: 2024-06-06
교환처: 스타벅스
상품명: 아이스카페아메리카노T+7레이어가나슈케이크

<그림 2> 확률적 예외기반 텍스트 제거 알고리즘의 적용 결과를 시각화한 웹 애플리케이션

Test Log	
NAME	Log
SortedArray_Probabilistic	버퍼링 디스크-1-1
	이벤트 참여확률서 고정위 호호
	이OPC TUNN-1 이벤트 지배하다-
	KING
	버거킹
	버거킹 와퍼세트
Probabilistic	340
	4
	0
	0
	0
	0
	0
	0
	0
	0
	0
Probabilistic Selector: -> null 해당 문자열은 존재하지 않습니다 최종배열: 버거킹,버거킹 와퍼세트 Exchange_Location:버거킹 유효기간: 2023-10-13 교환처: 버거킹 상품명: 와퍼세트	

Test Log	
NAME	Log
Probabilistic	87380
	12
	8
	0
	0
	0
Indices_Probabilistic	0
	1
	3
	4
	2
	5
ProbabilisticSelector	6
	0
	1
Probabilistic Selector: -> null 해당 문자열은 존재하지 않습니다 최종배열: 버거킹,버거킹 와퍼세트 Exchange_Location:버거킹 유효기간: 2023-10-13 교환처: 버거킹 상품명: 와퍼세트	

(a) 확률적 예외기반 텍스트 제거 알고리즘 테스트(1)

(b) 확률적 예외기반 텍스트 제거 알고리즘 테스트(2)

<그림 3> 확률적 예외기반 텍스트 제거 알고리즘의 적용 결과를 시각화한 웹 애플리케이션 테스트 로그

알고리즘의 적용 결과를 시각화한 웹 애플리케이션 테스트 화면이다[9]. 이 화면은 각 Release의 예외 처리 단어가 포함되었는지 확인하고, 이를 바탕으로 연속 일치 가중치 계산 공식인 수식(1)을 사용하여 확률값을 계산한다. 이렇게 계산된 확률값을 기준으로 대조군 데이터를 재정렬하며, 이는 원래 데이터의 실제 인덱스 값과 함께 출력된다. 또한 특정 임계값을 초과한 경우 해당 데이터는 확률적 선택자로 판단되어 별도로 보여준다. 입력된 모바일 교환권 이미지는 이전에 설명한 확률적 예외기반 텍스트 제거 알고리즘의 프로세스를 통해 분석된다. 이 과정을 통해 모바일 교환권의 교환처, 사용 유효기간, 그리고 교환 가능한 상품 정보를 자동으로 파악하고 추출할 수 있음을 보여준다.

V. 결론

참고문헌

- [1] 안세진, 황현호, 임진희, “종이기록 데이터화를 위한 AI-OCR 적용 사례연구,” 정보관리학회지 제39권, 제3호, 2022, pp. 165-193.
 - [2] 이승우, 김성진, 윤영현, 백재순, “안드로이드 환경에서 화면 캡처와 OCR을 활용한 실시간 번역 애플리케이션 개발,” 한국컴퓨터정보학회 학술발표논문집, 2023, pp. 267-268.
 - [3] 박혜지, 이상훈, 이영학, “OCR을 이용한 자동차 번호판 인식 시스템,” Proceedings of KIIT Conference, 2022, pp. 430-434.
 - [4] Naver Cloud, “Clova OCR,” last updated May 25, 2023. [Online].
<https://guide.ncloud-docs.com/docs/ko/clovaocr-overview>
 - [5] 박원익, “NAVER's AI text recognition team breaks world record,” Chosun Biz, 12-Jun-2018. [Online].
https://biz.chosun.com/site/data/html_dir/2018/06/12/2018061202617.html.
 - [6] 이소현, “온라인 선물하기 시장의 주요 트렌드 변화,” 소비자 평가, 06-Feb-2023. [Online].
<http://www.iconsumer.or.kr/news/articleView.html?idxno=25451>.
 - [7] QuaggaJS, “An advanced barcode-scanner written in JavaScript,” GitHub repository, 2021. [Online].
<https://github.com/ericblade/quagga2>.
 - [8] JsBarcode, “Barcode generation library written in JavaScript,” GitHub repository, 2021. [Online].
<https://github.com/lindell/JsBarcode>.
 - [9] 김찬술, “확률적 예외기반 텍스트 제거 알고리즘의 적용 결과를 시각화한 웹 애플리케이션,” Coupongo, 14-Nov-2023 [Online].
<http://http://coupongo.co.kr/>.
-