잔류물질정보 데이터베이스 자동 업데이트 시스템

신무곤, 백의준, 박지태, 김명섭
고려대학교

{tm0309, pb1069, pjj5846, tmskim}@korea.ac.kr

Automatic Database Update system for Pesticides and Veterinary Drugs Information

Mu-Gon Shin, Ui-Jun Baek, Jee-Tae Park, Myung-Sup Kim
Korea Univ.

요 약

현재 잔류물질정보를 제공하는 웹 페이지에서는 농약 및 동물용 의약품의 식품 내 잔류허용기준 정보를 제공하고 있다. 하지만 정보의 누락, 검색 시간 지연, 웹 페이지 오류 등 정보 제공이 원활하게 이루어지지 않고 있기에 사용자들이 불편을 겪고 있다. 또한 잔류허용기준이 개정 될 때마다 수동으로 정보를 업데이트 해야하는 등 불편함이 있다. 이에 본 논문에서는 잔류물질정보 제공 웹 페이지의 데이터베이스 개선방안과 잔류허용기준 개정고시 문서 및 기타 데이터의 자동 업데이트 시스템 구조에 대해 설명한다.

Ⅰ. 서 론

잔류물질정보를 제공하는 웹 페이지 에서는 농약 및 동물용 의약품의 식품 내 잔류허용기준 정보와 농약, 의약품 정보, 농약의 분석법 등 정보를 제공하고 있다[1]. 이러한 정보를 통해 국내 기업 및 농민들은 자신의 생산품의 식품 내 잔류허용기준을 측정하고 그 측정값이 국내 혹은 국제 기준을 통과할 수 있는지 가늠하는 척도로 사용하고 있다. 그러나 이 페이지를 그대로 사용하기에는 정보 누락, 검색 시간 지연, 웹 페이지 오류, 검색 불가 등 많은 문제들이 존재한다.

먼저, 농약 정보를 검색 하였을 때 전체 정보 중 약 74%의 공백이 있고 해외 자료를 제외하고는 약 60%의 정보가 누락되어 있는 것을 확인하였다. 또한 가장 중요한 정보라고 할 수 있는 국내 농약 잔류허용기준의 경우에는 약 72% 공백이 발견되어 심각한 문제를 가지고 있는 것으로 파악된다. 동물용 의약품의 경우도 마찬가지로, 많은 데이터의 누락이 있어 실제 이용자들이 효과적인 정보를 얻어가는지는 미지수이다.

그리고 정보 검색 시 많은 양의 정보를 로드 하지 않음에도 지연이 발생하며 농약 명으로 검색 시에는 농약 명이 일치함에도 관련 자료가 검색 되지 않는 등 웹 페이지 내의 오류가 빈번히 발생하고 있다. 또한 잔류허용기준이 개정될 때마다 수동으로 업데이트된 정보를 입력해야 하기 때문에 정보의 수정에 많은 어려움이 있다.

이러한 문제점들이 많이 존재하기 때문에 잔류물질정보 데이터베이스를 최신화 하고 데이터의 공백을 채우는 것이 중요하다. 본 논문에서는 데이터베이스를 최신화 하고 데이터의 공백을 채우기 위해 개정고시 문서 및 데이터 자동 업데이트 시스템을 제안한다.

본 논문은 서론에서 연구 배경과 목표를 서술하고, 본론에서 자동업데이트 시스템 구조를 제안한다.

본 연구는 2020 년도 식품의약품안전처의 연구개발비 (20162 수산물 625)로 수행된 연구임..

Ⅱ. 본론

본 장에서는 잔류물질정보 웹 페이지의 문제점을 서술하고 개정고시 문서 및 데이터 자동 업데이트 시스템 구조에 대해 설명한다.

2.1 웹 페이지의 문제점

기존의 잔류물질정보 웹 페이지의 데이터는 수동으로 업데이트 되고 있다. 잔류허용기준은 개정이 빈번하게 일어나고 한번에 많은 양이 업데이트 된다. 이를 수동으로 수정하고 입력하는 것은 많은 시간이 걸리는 작업이기 때문에 이를 자동화하는 것은 매우 중요하다.

농약 명으로 검색 시 농약 명이 일치함에도 관련 자료가 검색되지 않는 웹 페이지의 오류가 발생하고 있다. 또한 검색 시 많은 정보를 로드 하지 않음에도 지연이 길게 발생하는데, 이는 데이터베이스의 색인화가 되어 있지 않거나 웹 서버에서 데이터베이스에 요청하는 쿼리에 문제가 있는 것으로 보여진다. 이를 개선함으로써 이용자들의 편의성을 증대시킬 수 있다.

그림 1. 농약 명 검색 시 오류

	공전번호	농약명	농약이명	시약명	잔류물의정의	용도	적용대상작물	계통	IUPAC명	분자식
빈칸개수	1301	0	1069	17	1747	619	1382	1330	1134	519
공백 비율(%)	72.56	0	59.62	0.95	97.43	34.52	77.08	74.18	63.25	28.95
	상품명	구조식	분석법	분석법해설서	형태	녹는점	끓는점	증기압	LogP_ow	밀도
빈칸개수	518	833	0	1258	869	926	1781	936	1113	1778
공백 비율(%)	28.89	46.46	0	70.16	48.47	51.65	99.33	52.2	62.07	99.16
	pk_a	용해도	안전성	국내 농약잔류허용기준		ADI(한국)	Acute RfD(한국)	RESIDUE(한국)	독성요약정보(한국)	관련파일(한국)
빈칸개수	1540	860	990	1308		1267	1593	1793	1792	1418
공백 비율(%)	85.89	47.96	55.21	72.95		70.66	88.85	100	99.94	79.09
	ADI(CODEX)	Acute RfD(CODEX)	RESIDUE(CODEX)	독성요약정보(CODEX)	관련파일(CODEX)	ADI(미국)	Acute RfD(미국)	RESIDUE(미국)	독성요약정보(미국)	관련파일(미국)
빈칸개수	1548	1648	1630	1792	1616	1644	1755	1793	1792	1734
공백 비율(%)	86.34	91.91	90.91	99.94	90.13	91.69	97.88	100	99.94	96.71
	ADI(일본)	Acute RfD(일본)	RESIDUE(일본)	독성요약정보(일본)	관련파일(일본)	ADI(유럽)	Acute RfD(유럽)	RESIDUE(유럽)	독성요약정보(유럽)	관련파일(유럽)
빈칸개수	1457	1786	1793	1792	1647	1495	1613	1792	1792	1589
공백 비율(%)	81.26	99.61	100	99.94	91.86	83.38	89.96	99.94	99.94	88.62
전체공백비율	**74.44**									
해외자료 제외	**60.95**									

그림 2. 잔류물질정보 데이터 현황

그리고 농약 및 동물용 의약품 정보를 검색 하였을 때 정보의 누락이 많이 발견된다. 누락 정보에는 해외의 잔류허용기준, 약품의 특성, 분석법 등 이용자들이 많이 찾을 법한 정보가 포함되어 있다. 본 연구팀은 전체 정보 중 약 74%, 해외 자료를 제외하고는 약 60%의 정보 누락을 발견하였다.

2.2 개정고시 문서 자동 업데이트 시스템

본 절에서는 앞서 언급한 웹 페이지의 문제점 중 한가지인 데이터 수동 업데이트를 개선하기 위해 잔류허용기준 개정고시 문서 자동 업데이트 시스템을 제안한다. 제안된 시스템을 통해 관리자는 편리하게 업데이트된 정보를 데이터베이스에 업데이트 할 수 있게 되고 사용자는 업데이트 된 정보를 파악하지 못하여 피해를 보는 일이 없어질 것이다.

먼저 개정고시 문서는 hwp 파일로 작성되기 때문에 파싱하여 데이터를 추출하기가 어렵다. 따라서 hwp 문서를 html 문서로 변환하는 과정이 필요하다. html 변환을 수행하기 전에 프로그램 상에서 예외처리가 힘든 부분을 제거하는 작업이 필요하다. 변환된 html 문서에서 필요한 데이터만 추출하기 위하여 전처리 과정이 필요한데, 이 과정에서는 추출 할 데이터 이외의 태그를 제거한다.

다음 과정은 텍스트 추출 단계이다. 이 단계에서는 테이블을 제외한 키워드들을 추출한다. 추출하는 키워드들은 그림 3 과 같다.

(1) 겐타마이신(Gentamicin) : 항균제

◎ 잔류물의 정의 : Gentamicin C1a, C2, C2a 및 C1의 합을 gentamicin으로 함

그림 3. 추출 키워드

이 단계를 거쳐 추출된 텍스트 중 필요하지 않은 텍스트를 제거한다. 문자열 검사를 통해 텍스트를 제거하는데 제거하는 텍스트는 다음과 같다.

1. (숫자), ◎ 글머리
2. “잔류물의 정의”

이렇게 추출된 텍스트들을 리스트의 형태로 병합하고, 중복을 제거한다. 농약(동물용의약품)의 용도 같은 경우 문서마다 존재하는 문서가 있고 아닌 문서가 있기 때문에 리스트의 형태는 달라질 수 있다. 텍스트 리스트의 형태는 다음과 같다.

[농약명, (용도), 정의]

다음 단계는 테이블 추출 단계이다. 이 단계에서는 테이블 데이터를 추출하여 리스트로 병합하는 작업을 수행한다. 여기에서 추출하는 데이터는 그림 4 와 같다

식품명	mg/kg	식품명	mg/kg	식품명	mg/kg
소근육	0.1	돼지간	2.0	가금신장	0.1
소간	2.0	돼지지방	0.1	가금지방	0.1
소지방	0.1	돼지신장	5.0	유	0.2
소신장	5.0	가금근육	0.1	넙치	0.1
돼지근육	0.1	가금간	0.1	송어	0.1
잉어	0.1	양근육	0.1	말근육	0.1
염소근육	0.1	양간	2.0	말간	2.0
염소간	2.0	양지방	0.1	말지방	0.1
염소지방	0.1	양신장	5.0	말신장	5.0
염소신장	5.0				

그림 4. 테이블 데이터

테이블에서 추출된 데이터를 재정렬하여 dictionary 형태의 배열로 저장한다.

여기까지 추출된 텍스트, 테이블 데이터를 병합하는 작업이 필요하다. 인덱스 별 검사를 수행하여 테이블 데이터를 텍스트 리스트에 병합한다. 병합된 최종 리스트의 형태는 다음과 같다.

[농약명, (용도), 정의, {적용식품:잔류허용기준} list]

Ⅲ. 결론

본 논문에서는 잔류물질정보 웹 페이지의 개선을 위해 개정고시 문서 자동업데이트 시스템을 제안하였다. 제안된 시스템을 통해 수작업으로 진행하던 개정고시 업데이트를 자동으로 수행할 수 있어 작업의 효율이 증대 될 것으로 기대된다.

향후 연구로는 공백을 채우기 위한 데이터 수집, 데이터베이스 개선을 위한 효율적인 쿼리 작성을 진행할 예정이다.

참 고 문 헌

[1] “식품안전나라.” 잔류물질정보.
https://www.foodsafetykorea.go.kr/residue/main.do