단어 유사도를 활용한 이용약관 내 개인정보 수집 항목 시각화 서비스

Visualization Service for Personal Data Collection in Terms of Service Using Word Similarities

요 약

본 연구는 온라인 서비스 이용 시 복잡한 이용약관을 사용자가 쉽게 이해할 수 있도록 크롬 플러그인 형태의 Privacy-Bot 서비스를 개발하였다. Privacy-Bot은 사용자가 요청하지 않아도 이용약관 페이지에 접속하면 자동으로 개인정보 수집 내역을 분석하며, 단어 유사도 분석을 통해 다양한 표현으로 명시된 개인정보 항목을 정확하게 감지한다. 분석 결과는 시각적으로 제공되어 사용자가 쉽게 파악할 수 있도록 구현되었으며, 사용자 스터디를 통해 Privacy-Bot의 개인정보 인식 효과와 활용 가능성을 확인하였다.

1. 서 론

온라인 서비스 이용 시, 이용약관은 사용자와 서비스 제공자 간의 권리와 의무를 명확히 하고, 법적 보호를 제공하는 중요한 문서이다. 약관을 제대로 읽지 않을 경우, 사용자는 서비스 이용 조건이나 개인정보 처리 방식을 제대로 이해하지 못해서 예기치 않은 문제나 법적 분쟁에 휘말릴 수 있다. 그럼에도 불구하고, 온라인 매체를 사용하는 이용자들이 이용약관 페이지를 마주했을 때 이용약관을 읽지 않고 동의하는 비율이 무려 70%에 달한다[1]. 이용약관을 읽지 않는 가장 큰 이유는 이용자들이 이용약관을 읽기에는 너무 길고 복잡하기 때문이다[2].

온라인 서비스 사용자가 이용약관을 제대로 읽지 않고 개인정보 수집에 동의하면 예기치 못한 정보가 업체에 넘어갈 수 있다. 예를 들어, 이용약관에 사용자의 다양한 개인정보를 수집하고 이를 광고 또는 인공지능 학습에 활용한다는 내용이 있더라도, 이용약관을 제대로 읽지 않은 사용자는 이를 모른 채로 동의하고 가입하게 될 수 있다. 나중에 이 사실을 알게 되어 개인정보 유출 등의 불이익이 생기더라도, 해당 내용에 이미 동의했기 때문에 법적 분쟁에서 불리하게 작용하게 된다.

이런 문제들을 해결하기 위해 사용자들이 이용약관을 빠르고 쉽게 이해할 수 있도록 하기위한 다양한 연구들이 수행되었다[3,4,5,6]. 하지만 기존 연구에서는 사용자가 필요할 때 마다 직접 개인정보 수집 분석을 요청해야 하거나, 해당 웹사이트 개인정보 수집 내역 분석을 위한 규약을 잘 지키고 있지 않으면 분석이 어렵다는 한계가 존재한다.

본 연구에서는 이러한 한계를 극복하기 위해 크롬 플러그인 형태의 Privacy-Bot 서비스를 개발하였다. Privacy-Bot은 사용자가 요청하지 않아도 이용약관 페이지에 접속하면 자동으로 개인정보 수집 내역을 분석한다. 단어 유사도 분석 모델을 활용해서 각 이용약관에서 개인정보를 다르게 명시하더라도(예: 전화번호, 전화, 모바일) 이를 정확히 감지할 수 있으며, 분석이 완료되면 사용자가 이해하기 쉬운 방식으로 시각화 하여 즉각적으로 개인정보 수집 내역을 확인할 수 있다. Privacy-Bot을 활용한 사용자 스터디를 수행해서, 사용자의 이용약관 페이지 접속 시 개인정보 수집 내역에 대한 인식 효과를 높일 수 있음을 확인했다.

2. 관련 연구

사용자가 복잡한 이용약관을 쉽게 이해할 수 있도록 돕기 위한 다양한 연구가 수행되었다. 예를 들어, Zimmeck & Bellovin[3]은 사용자가 특정 웹사이트의 개인 정보 이용약관 분석을 요청하면, 해당 웹사이트의 개인정보 수집 여부, 보관 기간, 광고주에게 공개 여부 등을 시각적으로 보여주는 시스템을 개발했다. Harkous et al.[4]은 사용자가 특정 웹사이트 URL을 입력하면, 해당 사이트의 개인정보 수집과 관련된 질문에 답변해주는 AI 챗봇을 개발했다. Agulo et al.[5]은 웹사이트에서 사용자의 행동 로그 데이터를 분석하여,

업체에 제공되는 사용자의 개인 정보(GPS 위치 정보, 주소 정보 등)를 시각화 하는 서비스 프로토타입을 제안했다. 하지만 이러한 연구들은 사용자가 직접 이용약관 분석을 요청해야 한다는 한계가 있다.

사용자의 별도 요청이 없더라도 수집되는 개인정보를 분석해서 알려주는 연구도 존재한다. Vu et al.[6]는 사용자가 사전에 정의한 개인정보 항목과 웹사이트의 개인정보 정책을 비교하여, 수집되는 개인 정보 항목을 시각화 해주는 프로그램인 Privacy Bird를 개발했다. 하지만 Privacy Bird는 웹사이트가 사용자의 개인정보 보호정책을 표준화된 형식으로 게시할 수 있도록 개발된 P3P(Platform for Privacy Preferences)[1] 기술이 적용된 웹사이트에만 사용할 수 있다는 한계가 있다.

본 연구에서는 기존 연구의 한계를 극복하고자, 사용자가 별도로 개인정보 수집 분석을 요청하지 않아도 자동으로 분석을 수행하고, 다양한 표현 방식으로 명시된 개인정보 항목을 감지해낼 수 있는 크롬 플러그인 형태의 Privacy-Bot 서비스를 제안한다. 이를 통해, 사용자가 웹서핑 도중에도 이용약관 페이지에 접속할 때마다 별도의 추가 작업 없이 시각화 된 이용약관 개인정보 수집 내역을 확인 할 수 있다.

3. Privacy-Bot 아키텍처

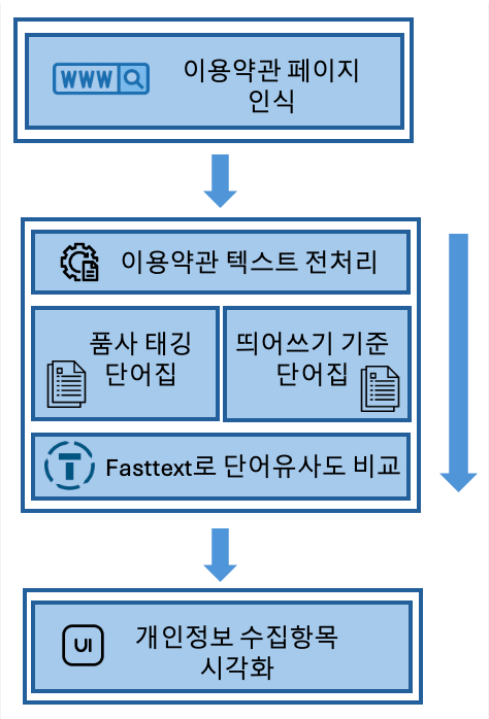

그림 1. Privacy-Bot 아키텍처

Privacy-Bot의 아키텍처는 [그림 1]과 같다. 사용자가 웹서핑 중 이용약관 페이지에 접속하면 이를 자동으로 감지하고 개인정보 수집 항목 분석을 수행한다. 단어 유사도 모델을 활용해서, 각기 다른 용어로 명시된 개인정보 항목도 정확히 감지하고 이를 팝업창으로 띄워서 사용자가 쉽고 빠르게 개인정보 수집항목을 인식할 수 있도록 구현되었다.

3.1 이용약관 페이지 인식

사용자가 웹서핑을 할 때, 방문한 페이지의 URI Path에 사전 지정된 단어(“Sign”, “Join”, “Member”)가 등장하면, Privacy-Bot은 해당 페이지가 회원가입 이용약관을 의미하는 페이지로 판단한다. 해당 페이지가 회원가입 이용약관 페이지로 판단되면 해당 웹페이지의 모든 텍스트가 서버로 전달된다. 서버는 트위터 형태소 분석기(OKT: Open Korean Text)[2]를 활용해서 전달받은 텍스트를 전처리한다. 먼저 불용어('의', '이다', '있다', '는', '다' 등)를 제거하고, 이용약관 검사를 위한 단어 리스트를 생성한다. 단어 리스트를 생성하기 위해, 먼저 품사 태깅을 통해 이용약관에서 명사 리스트를 만든다. 예를 들어, “휴대폰번호를 수집한다" 는 문장에서 [“휴대폰”, “번호”, “수집"]를 추출한다. 추가로, 띄어쓰기를 기준으로 이용약관에서 단어를 추출한다. 예를 들어, “휴대폰번호를 수집한다"는 문장에서 [“휴대폰번호”, “수집"]를 추출한다. 이 두 방법으로 추출한 단어를 합쳐서, 최종적으로 [“휴대폰”, “번호”, “휴대폰번호”, “수집”]이라는 단어 리스트를 생성할 수 있다. 이 단어 리스트에서 단어 유사도 비교를 통해 어떤 개인 정보가 수집되는지 파악할 수 있다.

3.2 개인정보 수집항목 파악

사전에 회원가입시 주로 수집되는 개인정보 11개 항목에 대한 텍스트("아이디", "비밀번호", "이메일", "이름", "생년월일", "전화번호", "아이피", "위치정보", "접속기록", "사진", "기기정보")를 선정했다. 이 11개 개인정보 항목을 해당 웹사이트에서 수집하는지 확인하기 위해, 위에서 생성한 단어 리스트를 활용한다.

단어 임베딩을 학습한 자연어 처리 모델인 Fasttext 모델[7]을 사용해서, 11개 개인정보 항목과 해당 웹사이트의 단어 리스트의 유사도를 계산했다. 11개 개인정보 항목과 유사도가 0.5 이상인 단어가 등장하면 이용약관이 해당 개인정보를 수집하는 것으로 판단했다. 예를 들어, 본 연구에서 정의한 개인정보 수집 항목 중 하나인 “전화번호”가 해당 페이지에서 수집되는지 확인하기 위해서 해당 페이지의 모든 단어 리스트를 검사한다. 만약 “전화번호”라는 단어가 해당 페이지의 단어 리스트에 없더라도 “휴대전화번호”(유사도 0.527), “휴대폰번호”(유사도 0.505)의 단어가 존재한다면 “전화번호”항목을 수집하는 것으로 인식한다.

[1] https://www.w3.org/P3P/

[2] https://github.com/twitter/twitter-korean-text

3.3 개인정보 수집항목 시각화

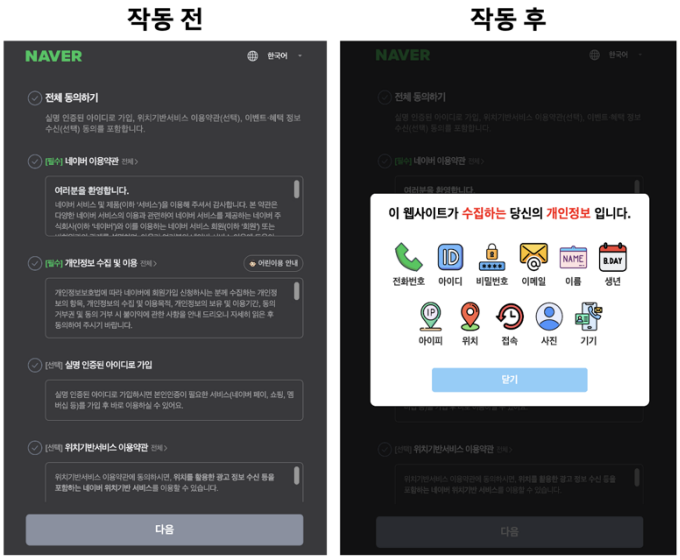

그림 2. Privacy-Bot 동작 화면 (네이버 회원 가입 페이지)

유사도 분석을 통해, 해당 페이지의 이용약관에서 사용자의 어떤 개인정보를 수집하는지 파악이 되는 즉시 Privacy-Bot은 11개의 개인정보 중에서 실제로 수집되는 항목을 팝업으로 표시해준다 [그림 2].

4. 사용자 스터디

본 연구에서 제안한 Privacy-Bot의 사용성을 검증하기 위해 20대 참여자 10명(남 8명, 여 2명)을 모집하여 사용자 스터디를 수행했다. 참여자를 5명씩 두 그룹으로 나누고, 한 그룹은 Privacy-Bot 크롬 플러그인을 활성화한 상태에서, 다른 그룹은 Privacy-Bot 크롬 플러그인을 활성화하지 않은 상태에서 크롬 브라우저로 특정 웹사이트에서 회원가입 하는 실험을 진행했다.

국내 웹사이트 중 방문자 수가 상위 10위 [3] 내에 위치하면서도 10명의 모든 참여자가 회원가입 하지 않은 온라인 커뮤니티인 "에펨코리아 [4] "를 실험 대상 웹사이트로 선정하고, 본 연구에서 정의한 11가지 개인정보 중 6가지 개인정보(IP주소, 이메일, 사용자ID, 비밀번호, 접속기록, 기기정보)를 수집하는 것을 미리 확인했다. 각 참여자는 본 연구진이 준비한 컴퓨터로 "에펨코리아"에 접속해서 회원가입을 수행했다. Privacy-Bot 크롬 플러그인이 활성화 된 그룹의 참여자에게는 Privacy-Bot의 에펨코리아 회원가입 이용약관에서 수집하는 개인 정보가 자동으로 안내되었다. 실험 진행자는 참여자가 개인 정보 입력을 마치고 회원가입 완료 버튼을 누르려고 할 때, "해당 웹사이트의 이용약관에서 어떤 개인정보를 수집하는지 기억하십니까?" 라고 질문했다. 이를 통해 각 참여자가 이용약관의 개인정보 수집 내역에 대해 얼마나 인지하고 있는지 확인했다.

실험 결과, Privacy-Bot 크롬 플러그인을 활성화하지 않고 회원가입을 수행한 그룹은 해당 웹사이트에서 수집하는 개인 정보가 무엇인지 전혀 답하지 못했다. 반면에, Privacy-Bot 크롬 플러그인을 활성화하고 회원가입을 수행한 그룹에서는 수집하는 개인 정보 중 평균 1.6개(P1: IP주소/이메일, P2:IP주소/이메일, P3:IP주소, P4:IP주소/이메일/사용자ID, P5:기억못함)를 대답했다. 이를 통해 Privacy-Bot이 이용약관의 개인정보 수집 사실을 사용자들에게 인식 시켜주는데 매우 효과적일 수 있음을 확인했다.

5. 결론

본 연구에서는 기존 이용약관 분석 시스템의 한계를 극복하고, 사용자가 별도로 요청하지 않아도 자동으로 개인정보 수집 내역을 분석 및 시각화 하는 Privacy-Bot 서비스를 개발하였다. 단어 유사도 분석을 통해 다양한 표현으로 명시된 개인정보 항목을 정확히 감지하고, 사용자가 쉽게 이해할 수 있도록 시각적으로 제공함으로써, 개인정보 수집 내역에 대한 인식 효과를 높였다. 사용자 스터디를 통해, 본 연구가 사용자들에게 이용약관이 수집하는 개인정보를 인식시키는데 효과적인 것을 입증했다. 향후연구로, 수집되는 개인정보 항목을 추가하고 단어 유사도 분석 모델을 개선하면 사용자의 개인정보 인식을 강화할 수 있는 유용한 도구로 활용될 수 있을 것으로 기대된다.

6. 참고논문

[1]백봉삼, "인터넷 이용약관 동의, 10명 중 7명 '안 보고 한다'", ZDNet Korea, 2020.11.23

[2]McDonald & Cranor. "The Cost of Reading Privacy Policies", A Journal of Law and Policy for the Information Society, vol.4, no.3, 543-568, 2008

[3] Zimmeck & Bellovin. "Privee: An Architecture for Automatically Analyzing Web Privacy Policies", 23rd USENIX Security Symposium, 1-16, 2014

[4] Harkous et al. "Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning", 27rd USENIX Security Symposium, 531-548, 2018

[5] Agulo et al. "Usable Transparency with the Data Track: A Tool for Visualizing Data Disclosures", CHI EA '15: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, 1803-1808, 2015

[6] Vu et al. "Influence of the Privacy Bird® user agent on user trust of different web sites", Computers in Industry, 311-317, 2010

[7] Bojanowski et al. "Enriching Word Vectors with Subword Information", Transactions of the Association for Computational Linguistics, 135-146, 2017

[3] https://ko.semrush.com/
[4] https://www.fmkorea.com/