개인화된 검색 결과 제공을 위한 머신러닝 알고리즘의 적용

김가원 김수찬 김지민 왕우석 윤제현 황선준

* 이산구조 - G팀

{kgw22, 8216kimsc, eric040222}@yonsei.ac.kr dhkddntjr001108@gmail.com

{jery0704, sunjun7559012}@yonsei.ac.kr

Application of machine learning algorithms to provide personalized search results

보고서팀 - 김지민 윤제현 황선준

I. 서론

우리 팀, G팀은 문제 1번을 선택했으며, 주제는 “개인화된 검색 결과 제공을 위한 머신러닝 알고리즘의 적용”을 주제로 선택했다. 이 주제는 이산구조에서 배운 알고리즘과 그래프 이론을 실제로 구현하고, 개인화된 검색 시스템의 중요성을 반영하기 위한 것이다.

문제 해결을 위한 계획과 결과는 다음과 같다. 먼 웹 크롤링을 이용하여 구글 검색 기록 데이터를 추출 후, 검색 기록과 유사한 내용을 나열하여 출력하며, 사용자가 입력한 데이터를 따로 파일에 저장한다. 마지막으로 크롤링을 통해 얻은 정보와 사용자가 입력한 데이터를 비교하여, 빈도가 높은 정보를 우선순위로 출력하는 알고리즘을 구현한다.

II-1. 본론 1

- 팀원이 조사한 내용과 이산구조 수업에서 언급한 내용 정리

1)그래프 이론

그래프 이론의 적용은 웹 크롤링 데이터의 구조를 이해하고 분석하는 데 중요한 역할을 한다. 웹 페이지와 하이퍼링크를 각각 그래프의 노드와 에지로 나타내어 전체 웹을 하나의 거대한 그래프로 표현할 수 있다. 이를 통해 데이터의 관계를 시각적으로 이해하고, 효율적인 데이터 탐색과 검색을 가능하게 한다. 예를 들어, 웹 크롤러는 웹 페이지를 탐색하면서 하이퍼링크를 따라가며 데이터를 수집[2]한다. 이러한 과정은 그래프 이론에서 노드와 에지로 표현될 수 있으며, 이를 통해 웹 페이지 간의 연결성을 분석할 수 있다.

그래프 이론은 실생활에서 다양한 문제를 해결하는 데 유용하게 사용된다. 네비게이션 시스템에서는 최단 경로를 찾기 위해 그래프 이론을 활용하여 도로 네트워크를 모델링하고, Dijkstra 알고리즘이나 A* 알고리즘을 사용하여 최적의 경로를 계산한다. 또, 소셜 네트워크 서비스에서는 친구 추천 시스템을 구현하기 위해 그래프 이론을 사용하여 사용자 간의 관계를 분석하고, 중심 노드를 찾거나 커뮤니티를 탐지하여 사용자에게 적절한 친구를 추천한다. 또한, 물건을 배송하는 데 있어서 순회 판매원 문제(TSP)는 그래프 이론을 통해 최적의 경로를 찾는 대표적인 예시이다. 이와 같이 그래프 이론은 복잡한 데이터 구조를 시각적으로 표현하고 분석하는 데 매우 유용한 도구이다.

불리언 검색의 적용

불리언 검색 알고리즘은 검색 쿼리를 보다 정교하게 파싱하고, 관련성이 높은 결과를 제공하는 데 중요한 역할을 한다. 불리언 검색은 사용자가 입력한 검색 쿼리를 파싱하여 키워드와 불리언 연산자를 식별하고, 이를 구조적으로 분석하여 검색 결과의 관련성을 높이는

1) 공은배 외 6명, 『Rosen의 이산수학』, 퍼스트북, 2023년 8월 10일, 727~730쪽

2) 왕태수 외 5명, 「효과적인 데이터 수집을 위한 웹 크롤러 개선 및 동적 프로세스 설계 및 구현」, 한국정보통신학회논문지, 2022년 11월, 1730쪽 2.1~7

방법[1]이다. 예를 들어, 사용자가 “머신러닝 AND 알고리즘”이라는 쿼리를 입력하면, 불리언 검색 알고리즘은 두 키워드가 모두 포함된 문서를 검색한다. 이러한 방식은 검색 결과의 정확도를 높이고, 사용자가 원하는 정보를 빠르게 찾을 수 있도록 도와준다.

불리언 검색의 주요 장점은 다양한 검색 조건을 조합하여 검색을 수행할 수 있다는 것이다. AND, OR, NOT 연산자를 사용하여 특정 조건에 맞는 문서를 필터링할 수 있으며, 이를 통해 검색 결과의 관련성을 높일 수 있다.[2] 예를 들어. “머신러닝 OR 딥러닝” 쿼리는 두 키워드 중 하나라도 포함된 문서를 검색하여, “머신러닝 NOT 딥러닝” 쿼리는 머신러닝의 키워드는 포함하되 딥러닝은 키워드는 포함하지 않는 문서를 검색한다. 이러한 불리언 연산자의 조합은 검색 엔진의 유연성을 높이고, 다양한 사용자 요구에 맞는 검색 결과를 제공할 수 있게 한다.

자연어 처리 기술의 활용

자연어 처리(NLP) 기술은 검색 알고리즘의 성능을 향상시키는 데 중요한 역할을 한다. 자연어 처리 기술은 사용자의 입력을 보다 정확하게 이해하고, 이를 기반으로 최적의 검색 결과를 제공할 수 있도록 도와준다. 자연어 처리는 음성 인식과 텍스트 분석을 포함하여, 컴퓨터가 인간의 언어를 이해하고 처리하는 기술[3]이다.

자연어 처리의 핵심 과정은 두 가지로 나눌 수 있다. 자연어 이해(NLU)와 자연어 생성(NLG)이다. 자연어의 이해는 입력된 자연어의 정보를 토대로 컴퓨터가 의미를 해석하는 과정이며, 자연어 생성은 컴퓨터가 명령에 대한 응답을 자연어로 생성하는 과정이다. 예를 들어, 사용자가 “오늘 날씨가 어때?”라는 질문을 입력하면, 자연어 이해 과정을 통해 컴퓨터는 사용자가 날씨 정보를 알고 싶어 한다는 것을 이해하고, 자연어 생성 과정을 통해 “오늘의 날씨는 맑고, 기온은 20도입니다”와 같은 응답을 생성한다.

자연어 처리 기술은 검색 알고리즘이 사용자 요청에 더욱 정확하게 응답할 수 있도록 도와준다. 음성 인식 기술은 사용자의 음성 명령을 텍스트로 변환하고, 텍스트 분석 기술은 문장의 의미를 파악하여 적절한 검색 결과를 제공한다. 이를 통해 검색 엔진은 사용자 경험을 향상 시키고, 사용자 만족도를 높일 수 있다.

자연어 처리 기술 - 시맨틱 네트워크, 시소러스, 코퍼스 등

시맨틱 네트워크(semantic networks)는 단어들 간의 의미 관계를 그래프로 표현하여 컴퓨터가 단어의 의미를 이해할 수 있도록 돕는다. 이는 의미 유사성 계산, 정보 검색, 텍스트 요약 등에 유용하다. 시소러스(thesaurus)는 단어와 그 유의어, 반의어 등을 체계적으로 정리한 것이며, 워드넷(WordNet)은 단어 간의 의미 관계를 그래프로 표현하여 검색 및 자연어 처리에 사용된다. 예를 들어, "자동차"라는 단어의 유의어와 관련 단어들을 쉽게 찾을 수 있다. 코퍼스(corpus)는 실제 사용된 언어 자료를 모아 놓은 데이터베이스로, 자연어 처리 시스템의 학습 및 평가에 사용된다. British National Corpus(BNC) 같은 코퍼스는 언어 모델의 정확성과 신뢰성을 높이는 데 중요하다. 이처럼 자연어 처리 기술은 검색 알고리즘의 성능 향상, 사용자 경험 개선, 정보 검색 및 처리에 널리 활용되며, 그 발전은 앞으로도 지속될 것이다[4].

---

1) 김용, 주원균, 「태그결합을 이용한 불리언 검색에서 순위화된 검색결과를 제공하기 위한 시스템 설계 및 구현」, 2012년 12월 19일, 106~108쪽
2) 김용, 주원균, 「태그결합을 이용한 불리언 검색에서 순위화된 검색결과를 제공하기 위한 시스템 설계 및 구현」, 2012년 12월 19일, 108쪽
3) 맹혜련, 「인공지능 기반 자연어 처리 기술의 현황 및 서비스 연구: 3종류의 기계통번역 장치 정확도 분석」, 2018년, 13~37쪽
4) 공은배 외 6명, 『Rosen의 이산수학』, 퍼스트북, 2023년 8월 10일, 737쪽

휴리스틱 탐색의 응용

휴리스틱 탐색 알고리즘은 검색 과정에서 최적의 경로를 찾는 데 중요한 역할을 한다. 휴리스틱 탐색은 모든 경로를 일일이 탐색하지 않고, 경험적 지식 또는 직관에 기반한 휴리스틱 함수를 사용하여 유망한 경로를 우선적으로 탐색하는 방법이다. 이로 인해 탐색 공간이 줄어들어 문제 해결 속도가 빨라진다.

A* 알고리즘은 휴리스틱 탐색의 대표적인 방법 중 하나로, 최단 경로를 찾는 문제에서 많이 사용[1]된다. 이 알고리즘은 시작 노드에서 목표 노드까지의 최적의 경로를 찾기 위해 휴리스틱 함수를 사용하여 탐색을 효율적으로 진행한다. A* 알고리즘은 다익스트라 알고리즘과 그리디 베스트 퍼스트 탐색의 장점을 결합한 형태로 작동하며, 최적성과 안정성을 보장한다. 예를 들어, 네비세이션 시스템은 A* 알고리즘을 사용하여 도로 네트워크에서 최적의 경로를 찾고, 이를 통해 사용자가 목적지에 가장 빠르게 도착할 수 있도록 안내한다.

휴리스틱 탐색의 장점은 시간과 자원을 절약하면서도 높은 효율성을 유지할 수 있다는 것이다. 이를 통해 검색 엔진은 복잡한 데이터 탐색 문제를 효율적으로 해결하고, 사용자에게 빠르고 정확한 검색 결과를 제공할 수 있다.

정렬 알고리즘

정렬 알고리즘은 컴퓨터 과학에서 매우 중요한 문제 중 하나로, 다양한 응용 분야에서 널리 사용된다. 예를 들어, 전화번호부를 만들 때 이름을 알파벳 순서로 정렬하거나, 노래 목록 등을 정리할 때 노래 제목을 알파벳 순서로 정렬하는 경우 등이 있다. 정렬은 리스트의 원소를 특정 순서에 따라 배열하는 과정이다. 예를 들어, 리스트 [7, 2, 1, 4]를 정렬하면 [1, 2, 4, 7]이 되고, [d, h, c, a]를 정렬하면 [a, d, f, h]가 된다. 정렬 알고리즘의 효율성은 컴퓨팅 자원의 사용 비율에 큰 영향을 미치기 때문에 많은 연구가 이루어졌다. 정렬 알고리즘에 대한 연구가 활발한 이유는 여러 가지가 있다. 어떤 알고리즘은 구현이 쉬운 반면, 다른 알고리즘은 더 효율적이다. 우리가 정렬 알고리즘을 다루는 이유는 정렬이 중요한 문제이기도 하고, 이해와 구현을 통해 다양한 컴퓨터 과학의 개념들을 배울 수 있다.[2]

문자열 매칭 알고리즘

문자열 매칭 알고리즘은 특정 문자열 P(패턴)가 다른 문자열 T(텍스트) 내에 존재하는지 찾는 방법이다. 예를 들어, 패턴 “101”이 텍스트 “1001001”에 포함되어 있는지 찾는 것이다. 단순 문자열 매칭 알고리즘은 텍스트의 각 위치에서 패턴이 일치하는지 확인하는 방식으로 작동한다. 패턴이 일치하면 해당 위치를 출력한다. 이 알고리즘은 모든 위치에서 패턴을 검사하기 때문에 시간 복잡도는 $O((n-m+1)m)$이다. 단순하지만. 비효율적일 수 있다. 그러나 이해하기 쉽고 기본 개념을 익히는 데 유용하다.[3]

II-2. 본론 2

- 서론에서 언급한 문제와 주제 설명

우리 팀은, 문제 1번을 선택하였으며, 주제는 “개인화된 검색 결과 제공을 위한 머신러닝 알고리즘의 적용”을 주제로 선색하였다. 이 주제는 이산구조에서 배운 알고리즘과 그래프 이론을 실제로 구현하고, 개인화된 검색 시스템의 중요성을 반영하기 위한 것이다. 문제

1) 김명재, 정태충,「그라프에서의 휴리스틱 탐색에 관한 연구」, 1997년 9월 22일, 2478~2479쪽
2) 공은배 외 6명,『Rosen의 이산수학』, 퍼스트북, 2023년 8월 10일, 231~232쪽
3) 공은배 외 6명,『Rosen의 이산수학』, 퍼스트북, 2023년 8월 10일, 234쪽

해결을 위한 계획은 다음과 같다.

먼저 웹 크롤링을 이용하여 구글 검색 기록 데이터를 추출 후, 검색 기록과 유사한 내용을 나열하여 출력하며, 사용자가 입력한 데이터를 따로 파일에 저장한다. 마지막으로 크롤링을 통해 얻은 정보와 사용자가 입력한 데이터를 비교하여, 빈도가 높은 우선순위로 출력하는 알고리즘을 구현한다.

II-3. 본론 3

- 구현한 python 프로그램 설명

프로그램은 Python으로 작성되었으며, 다음과 같은 기능을 포함한다.

1. 파일 입출력 기능: 사용자가 입력한 키워드를 파일에 저장하고, 프로그램 실행 시 이전에 입력된 키워드를 불러온다
2. 웹 크롤링 기능: 사용자가 입력한 키워드를 기반으로 구글 검색 결과에서 추천 검색어와 해당 키워드가 포함된 문장을 수집한다.
3. 키워드 빈도 증가 및 필터링 기능: 사용자 입력 키워드의 빈도를 추적하고, 특정 키워드가 일정 횟수 이상 입력되면 우선순위로 설정한다.
4. 결과 출력 및 필터링 기능: 수집된 문장에서 단어를 추출하고, 특정 키워드가 포함된 단어만을 필터링하여 출력한다.

python 코드 QR코드

각 기능에서 팀원의 개념, 책 개념 적용

그래프 이론의 적용: 웹 크롤러는 웹페이지를 탐색하면서 하이퍼링크를 따라가며 데이터를 수집한다. 이러한 과정은 그래프 이론에서 노드와 에지로 표현될 수 있으며, 이를 통해 웹 페이지 간의 연결성을 분석할 수 있다.

불리언 검색의 적용: 사용자가 입력한 키워드를 기반으로 검색 쿼리를 파싱하고, 추천 검색어를 추출하여 관련성을 높인다. 불리언 연산자를 사용하여 특정 조건에 맞는 문서를 필터링할 수 있다.

자연어 처리 기술의 활용: 텍스트 분석을 통해 문장의 의미를 파악하고, 특정 키워드가 포함된 문장을 추출한다. 이를 통해 검색 결과의 정확도를 높이고, 사용자의 요구에 맞는 결과를 제공한다.

휴리스틱 탐색의 응용: 사용자가 입력한 키워드의 빈도를 추적하고, 특정 키워드가 일정 횟수 이상 입력되면 우선순위로 설정한다. 이를 통해 검색 결과의 효율성을 높이고, 사용자가 원하는 정보를 빠르게 찾을 수 있도록 한다.

문자열 매칭 알고리즘의 적용: 웹 크롤링을 통해 수집된 텍스트에서 사용자가 입력한 키워드가 포함된 문장을 찾아낸다. 이는 문자열 매칭 알고리즘을 사용하여 키워드가 포함된 문장을 효율적으로 추출하고, 거른다.

정렬 알고리즘의 적용: 필터링된 단어들을 정렬하여 출력할 때 사용된다. 사용자가 입력한 키워드가 포함된 단어들을 알파벳 순서로 정렬함으로써, 결과를 보기 쉽게 제공한다.

머신 러닝을 활용한 검색 알고리즘은 대량의 데이터를 수집하고 처리하며, 적절한 특성 엔지니어링과 모델 학습을 통해 성능을 최적화한다. 그래프 이론, 불리언 검색, 자연어 처리, 휴리스틱 탐색 등 다양한 알고리즘과 기술을 결합하여, 사용자에게 더 정확하고 개인화된 검색

결과를 제공할 수 있다. 이러한 기술은 검색 엔진의 사용성을 개선하고, 사용자 만족도를 높이는 데 크게 기여한다. 데이터 중심의 의사결정을 가능하게 하고, 실시간으로 최적화된 정보를 제공하여 검색 엔진의 효율성과 효과성을 극대화한다.

프로그램은 각 팀원이 제안한 개념을 활용하여 사용자에게 개인화된 검색 결과를 제공하는 구조로 설계되었다. 이러한 기술은 검색 엔진의 사용성을 개선하고, 사용자 만족도를 높이는 데 크게 기여한다. 데이터 중심의 의사결정을 가능하게 하고, 실시간으로 최적화된 정보를 제공하여 검색 엔진의 효율성과 효과성을 극대화한다.

III. 결론

우리 팀은 “개인화된 검색 결과 제공을 위한 머신러닝 알고리즘의 적용”을 주제로 프로젝트를 진행했다. 이 프로젝트에서는 그래프이론, 불리언 검색, 자연어 처리, 휴리스틱 탐색. 문자열 매칭, 그리고 정렬 알고리즘을 효과적으로 결합하여 사용자에게 보다 정확하고 개인화된 검색 결과를 제공하는 방법을 모색하였다.

먼저, 웹 크롤링을 통해 구글 검색 기록 데이터를 수집하고, 그래프 이론을 적용하여 데이터 간의 관계를 분석했다. 불리언 검색 알고리즘을 사용하여 검색 쿼리를 정교하게 파싱하고, 자연어 처리 기술을 통해 사용자의 입력을 정확하게 이해하여 관련성 높은 결과를 도출했다. 휴리스틱 탐색을 통해 검색 속도를 향상시키고, 사용자 경험을 개선하는 데 중점을 두었다.

결과적으로, 우리 팀이 개발한 시스템은 사용자 입력 키워드의 빈도를 추적하고, 가장 관련성 높은 정보를 우선순위로 제공함으로써 검색 정확도와 효율성을 크게 향상시켰다. 이를 통해 개인화된 검색 결과를 제공하는 시스템의 가능성을 확인할 수 있었으며, 다양한 알고리즘의 조합이 검색 엔진 성능에 긍적적인 영향을 미침을 입증했다.

IV. 후기

황선준: 이번 프로젝트는 많은 도전과 성장을 경험한 시간이었다. 코드 작성부터 보고서 작성까지 주도적으로 수행하며 알고리즘과 기술을 깊이 이해할 수 있었다. 이론을 실제 프로젝트에 적용하면서 이론과 실습의 간격을 줄일 수 있었다. 코드 구현 과정에서 문제를 해결하며 능력을 기를 수 있었고, 팀원들과의 협업으로 프로젝트 완성도를 높였다. 보고서 작성은 과정을 정리하고 배운 점을 되짚는 좋은 기회였다. 이번 경험은 앞으로의 학업과 연구에 큰 밑거름이 될 것이며, 더 큰 도전에 나아갈 자신감을 얻었다.

왕우석: 이번 팀 프로젝트는 저에게 많은 것을 깨닫게 해준 소중한 경험이었습니다. 처음에는 프로젝트라는 것을 협동하면 쉽게 된다고 생각했지만, 진행하면서 다양한 어려움에 부딪히게 되었습니다. 그럼에도 불구하고 팀원의 강점을 발휘하고 보완하며 협력한 결과, 개인이 만든 결과물보다 훨씬 뛰어난 성과를 얻을 수 있었습니다. 각자의 노력이 하나로 모여 프로젝트의 완성도를 높였고, 혁신적인 결과물을 만들어낼 수 있었습니다.

김수찬: 이전에는 검색엔진 알고리즘에 대해 잘 알지 못했고 대충 이럴 것이라는 상상만 했었지만 이번 기회로 검색엔진에 알고리즘이 어떻게 사용되는지 알 수 있는 좋은 기회가 되었던 것 같습니다.

김지민: 팀 프로젝트를 진행하면서 이산구조 시간에 배웠던 점을 이용하고 추가적인 자료조사를 하며 새롭게 알아간 점도 있었고 배웠던 것을 복습하는 계기가 되었습니다.역량이 부족하여 팀에 도움이 많이 못 되어 아쉽습니다.

윤제현: 평소에 관심있던 검색 알고리즘 주제에 대해서 조사할 수 있어서 유익했습니다. 실생활에서 쉽게 접할 수 있는 주제이지만 실제로 알고리즘에 대해서 자세히 조사해본 적이 없어서 아쉬웠습니다. 하지만 요번 팀플의 주제로 선정되어 그동안 몰랐던 검색 알고리즘의 원리에 대해서 알아보고 더 나아가 간단하게 직접 구현까지 해보는 유익한 시간이었습니다.
김가원: 이번 기회를 통해 검색 알고리즘이 어떠한 원리로 일상생활에 도움을 주는지 다소 심도있게 조사하면서 원리랑 종류 등을 알 수 있었습니다. 뿐만 아니라 팀 프로젝트를 통해 개개인의 자료가 모아지면서 각자 조사해온 분야에 대해 설명할 수 있고 다른 팀원들의 설명을 들으면서 다소 생소했던 분야에 대한 새로운 지식을 습득할 수 있는 유익한 시간이었습니다.