JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 6, pp. 115-127, June 2021
https://doi.org/10.9708/jksci.2021.26.06.115

Methodology for Search Intent-based Document Recommendation

Donghoon Lee*, Namgyu Kim*

*Graduate Student, Graduate School of Business IT, Kookmin University, Seoul, Korea
*Professor, Graduate School of Business IT, Kookmin University, Seoul, Korea

[Abstract]

It is not an easy task for a user to find the correct documents that a user really wanted at once from a vast amount of the search results. For this reason, various methods of recommending documents by taking the user's preferences into consideration based on the user's document browsing history have been proposed. However, the document recommendation methodology based on the document browsing history also has a limitation that only the information the user has viewed is utilized, but the intent of the user searching for the document is not fully utilized. Therefore, we propose a document recommendation method based on the user's search intent that utilizes information on "Why" the user reads the document, instead of the information on "Who" reads the document. In order to confirm the feasibility of the proposed methodology, an experiment was conducted by analyzing 239,438 actual user's search history of one of the most popular e-commerce platform companies in Korea. As a result, our methodology showed superior performance compared to the existing content-based or simple browsing history-based recommendation model.

▸ **Key words**: Document Recommendation, Search Intent, Text Mining, TF-IDF, User Access Log

[요 약]

방대한 데이터 가운데 사용자가 원하는 정보를 단번에 찾아내는 것은 결코 쉬운 일이 아니다. 이로 인해 사용자의 문서 열람 이력을 바탕으로 사용자 선호를 고려해 문서를 추천하는 다양한 방법들이 제안되었다. 하지만 기존에 활용된 문서 열람 이력 기반 문서 추천 방법론은 문서를 누가 열람했는지의 정보만을 활용할 뿐, 사용자가 해당 문서를 열람하게 된 의도(Intent)를 충분히 활용하지 못했다는 한계를 갖는다. 따라서 본 연구에서는 해당 문서를 누가(Who) 읽었는지의 정보가 아닌 해당 문서를 왜(Why) 읽었는지의 정보를 활용하는 검색 의도 기반 문서 추천 방안을 제시하고자 한다. 제안 방법론의 우수성을 확인하기 위해 국내 전자상거래 플랫폼 기업인 'C' 사의 실제 사용자 검색 이력 239,438건을 분석한 실험을 수행하였으며, 실험 결과 제안 방법론이 기존의 내용 기반 추천 모델 및 단순 열람 이력 기반 추천 모델에 비해 우수한 성능을 보임을 확인하였다.

▸ **주제어**: 문서 추천, 검색 의도, 텍스트 마이닝, TF_IDF, 사용자 접근 이력

• First Author: Donghoon Lee, Corresponding Author: Namgyu Kim
*Donghoon Lee (donghoonlee@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
*Namgyu Kim (ngkim@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
• Received: 2021. 05. 06, Revised: 2021. 06. 22, Accepted: 2021. 06. 24.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 정보기술의 발달에 따라 생성, 수집, 그리고 유통되는 데이터의 규모가 폭발적으로 증가하고 있다. IDC(International Data Corporation)는 최근 보고에서 2025년에 생산되는 데이터는 163제타바이트(Zettabyte)로 2016년을 기점으로 생산된 데이터양의 10배 수준이며, 이 중 90%가 비정형 데이터일 것으로 예상하였다[1]. 이렇듯 지속적인 데이터의 증가가 예상됨에 따라, 정보 소비자들이 양질의 정보에 손쉽게 접근할 수 있도록 지원하기 위한 방법에 많은 관심이 집중되고 있다. 대표적으로 구글(Google), 빙(Bing), 바이두(Baidu)와 같은 온라인 검색 포털 사이트들은 대량의 텍스트 및 이미지 데이터에 대한 사용자들의 정보 획득 비용을 절감시키기 위해 다양한 검색 기능을 제공하고 있다.

일반적으로 사용자들은 자신의 목적에 맞는 정보를 탐색하기 위해 온라인 포털 사이트에 접속하여 검색어를 입력하고, 검색 엔진은 입력받은 검색어에 대응되는 문서를 찾아서 사용자에게 보여준다. 하지만 이러한 방식을 통해 방대한 데이터 가운데 사용자가 원하는 정보를 단번에 찾아내는 것은 결코 쉬운 일이 아니다. 따라서 사용자들은 검색 결과에 대한 정렬 및 필터링을 통해 검색 범위를 좁히거나, 목적한 정보를 획득하지 못할 경우 다른 검색어로 재검색을 시도하는 등의 반복되는 탐색 과정을 거치게 된다. 이처럼 사용자가 방대한 데이터에서 양질의 정보를 획득하기 위해서는 많은 시간과 노력이 소비되며, 기하급수적으로 데이터가 증가하는 빅데이터 환경에서 사용자가 직접 정보를 탐색하는 방식은 사용자의 정보 획득 만족도를 저하시키는 요인으로 작용한다. 따라서 사용자의 정보 획득 만족도 향상을 위한 다양한 연구들이 꾸준히 수행되어 왔다.

전통적으로는 특정 정보와 내용이 유사한 정보를 관련 정보로 추천하는 방법들이 다양한 분야에서 널리 사용되고 있다. 예를 들어 특허 문서 간 유사도를 계산하여 유사 특허 문서를 제안한 연구[2], 정부 R&D 유사 과제를 분석한 연구[3], 법률 문서 분석을 위해 판결문 간 유사도를 활용한 연구[4] 등이 있으며, 언론 분야에서도 뉴스 기사 추천에 사용되는 주요 키워드를 도출하기 위해 문서의 유사도를 사용한 연구[5]가 수행되었다. 하지만 내용의 유사성에 기반하여 도출된 추가 정보는 사용자가 이미 탐색한 정보와 비슷한 내용을 갖는 경향이 있기 때문에, 사용자에게 중복된 정보를 반복적으로 제공한다는 한계, 그리고 사용자의 특성을 반영하지 못하는 한계를 가진다.

따라서 이러한 한계들을 극복하고자, 사용자의 문서 열람 이력 등 행동 패턴을 활용하여 문서를 추천하기 위한 연구들이 다수 수행되었다. 전통적으로는 문서 열람 이력이 유사한 다른 사용자들의 열람 정보를 활용하는 협업 필터링(Collaborative Filtering) 기법을 사용하여 뉴스, 영화, 음악, 상품 등을 추천하는 연구들[6-10]이 수행된 바 있으며, 두 사건이 동시에 발생하는 정도를 측정하는 연관 분석[11]을 사용하여 상품을 추천해주는 연구[12]도 진행되었다. 하지만 이러한 열람 이력 기반 문서 추천 방법론은 문서를 누가 열람했는지의 정보만을 활용할 뿐, 사용자가 해당 문서를 열람하게 된 의도(Intent)를 충분히 활용하지 못했다는 한계를 갖는다. 이러한 한계는 <Fig. 1>의 예를 통해 설명된다.

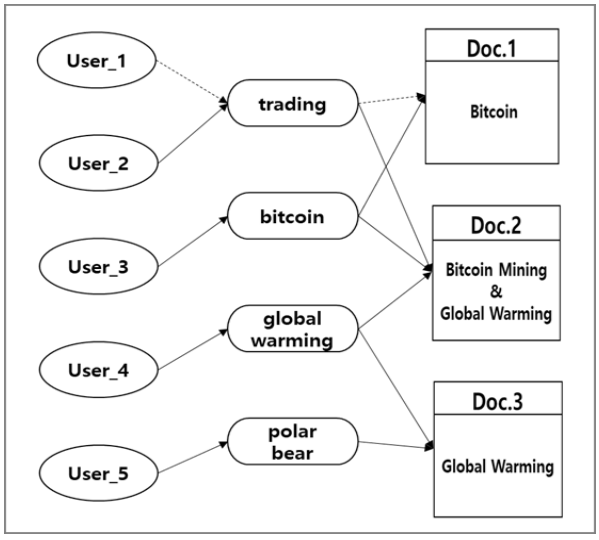


Fig. 1. Access History with Search Keywords

<Fig. 1>은 5명의 사용자가 3개의 문서 즉, 비트코인(Bitcoin) 관련 문서(Doc.1), 지구 온난화(Global Warming) 관련 문서(Doc.3), 그리고 비트코인 채굴과 지구 온난화 관련 문서(Doc.2)를 열람한 이력을 나타낸다. 또한 문서 획득에 사용된 4가지 검색어는 “trading”, “bitcoin”, “global warming”, 그리고 “polar bear”이다. User_1과 User_2에 연결된 점선과 실선은 User_1은 “trading” 검색어를 이용하여 Doc.1을 열람하였고, User_2는 동일한 검색어를 이용하여 Doc.2를 열람하였음을 구분하기 위해 사용되었다. 또한 User_3은 “bitcoin” 검색어를 통해 Doc.1과 Doc.2를, User_4는 “global warming” 검색어를 통해 Doc.2와 Doc.3을 열람하였다. 마지막으로 User_5는 “polar bear” 검색어를 통해 Doc.3만을 열람하였다. <Fig. 1>에 나타난 사용자

의 문서 열람 이력을 사용자 관점과 검색어 관점으로 나누어 파악한 결과가 <Fig. 2>에 제시되어 있다.

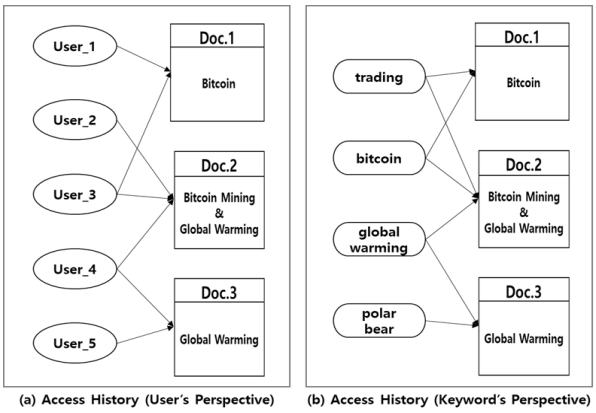


Fig. 2. Comparison of User's Perspective and Keyword's Perspective

<Fig. 2(a)>는 <Fig. 1>의 문서 열람 이력을 사용자 관점에서, 그리고 <Fig. 2(b)>는 본 연구에서 제안하고자 하는 방안으로 동일한 내용을 검색어 관점에서 파악한 것이다. 각 관점에서 Doc.2와 연관된 문서를 식별하는 과정은 다음과 같다. <Fig. 2(a)>의 경우 Doc.2와 Doc.1을 동시에 열람한 사용자 수와 Doc.2와 Doc.3을 동시에 열람한 사용자 수는 1명으로 서로 동일하다. 따라서 사용자 관점에서는 Doc.2의 연관 문서로 Doc.1과 Doc.3을 동일한 우선순위로 추천하게 된다. 한편 <Fig. 2(b)>의 경우 Doc.2와 Doc.1의 접근에 동시에 사용된 검색어 수는 2개이며, Doc.2와 Doc.3의 접근에 동시에 사용된 검색어 수는 1개이다. 따라서 검색어 관점에서 연관 문서를 추천하는 경우, Doc.2의 연관 문서로 Doc.3 보다 Doc.1을 높은 우선순위로 추천하게 된다.

<Fig. 2>를 통해 관점의 차이에 따라 동일한 문서 열람 이력에 대해서도 상이한 연관 문서를 추천하게 됨을 확인하였다. 사용자 관점 문서 추천의 경우 특정 문서를 누가(Who) 읽었는지에 집중하는 반면, 본 연구에서 제안하는 검색어 관점 문서 추천의 경우 사용자가 특정 문서를 왜(Why) 읽었는지에 집중한다. 본 연구에서는 사용자의 정보 탐색 의도가 누락된 기존 문서 추천의 한계를 극복하기 위해, 사용자의 문서 열람 의도가 담긴 검색어 정보를 활용하여 연관 문서를 식별하는 검색 의도 기반 문서 추천 방안을 제시하고자 한다.

본 논문의 이후 구성은 다음과 같다. 2장에서는 검색 및 추천 시스템, 그리고 텍스트 마이닝(Text Mining) 및 문서 유사도 분석에 대한 기존 연구를 소개한다. 3장에서는 본 연구에서 제안하는 문서 추천 방법론인 검색 의도 기반 문서 추천 방안의 개념 및 과정을 설명한다. 제안 방법론의 우수성을 확인하기 위해 국내 전자상거래 플랫폼 기업인 'C' 사의 실제 사용자 검색 이력 239,438건을 분석한 실험 결과는 4장에서 소개하고, 마지막 장인 5장에서는 본 연구의 기여와 한계, 그리고 향후 연구 방향을 제시한다.

II. Related Research

1. Information Retrieval and Recommendation Systems

정보 검색(Information Retrieval)이라는 용어는 매우 광범위한 의미로 사용되나, 본 연구 분야에서는 일반적으로 컴퓨터에 저장된 대량의 컬렉션에서 정보 요구를 만족하는 구조화되지 않은 자료를 찾는 행위를 의미한다[13]. 정보 검색을 수행하는 검색 엔진은 일반적으로 수집 부(Crawling Process), 색인 부(Indexing Process), 그리고 검색 및 순위화 부(Retrieval & Ranking Process) 등으로 구성된다. 세부적으로는 문서나 웹상의 데이터 등이 저장된 문서 집합을 색인(Indexing)하고 사용자의 정보 요구가 반영된 검색어를 처리한 후, 색인(Index)과 연관된 검색 결과를 순위화하여 상위 문서부터 사용자에게 보여주는 작업을 수행한다[14].

검색 엔진은 사용자의 정보 요구가 반영된 검색어와 검색 결과의 연관성을 결정하기 위한 일련의 과정을 수행한다. 하지만 검색 엔진의 검색 결과와 사용자의 정보 요구 간의 연관성을 측정하는 작업은 검색어에 담겨있는 사용자들의 정보 니즈(Need)를 파악해야 한다는 점에 매우 어려운 일로 평가받고 있다. 이러한 어려움은 빅데이터 시대로 진입하면서 이전보다 훨씬 다양하고 새로운 정보들이 웹상에 유통됨에 따라 더욱 부각되고 있으며, 이제는 사용자들이 어떤 정보를 알고 어떤 정보를 모르는지조차 인지하기 어려운 검색 환경에 직면하게 되었다. 따라서 이러한 한계를 극복하고자 사용자의 직접적인 정보 요구를 필요로 하지 않는 정보의 추천이 매우 중요한 요소로 부상하게 되었다.

추천 시스템의 종류는 매우 다양하지만[15], 본 연구에서는 사용자 간 이력 정보를 활용하는 관점과 아이템의 동시 발생 정보를 활용하는 관점에서 추천 시스템을 분류하여 소개한다. 먼저 동시 발생 정보를 활용하는 관점에서, 장바구니 분석이라고도 불리는 연관 분석은 동일한

트랜잭션(Transaction)에서 아이템이 동시에 출현하는 패턴을 연관 규칙으로 표현하고, 흥미성 척도들을 통해 가장 연관성이 높은 규칙을 찾아내는 방법이다. 흥미성 척도에 관한 연구도 이루어지고 있는데[16], 일반적으로 신뢰도(Confidence)와 지지도(Support)가 널리 사용되고 있다. 또한, 연관 분석을 추천에 적용하기 위한 다양한 연구[12, 17]들이 활발히 진행되고 있다.

다음으로 사용자의 이력 정보를 활용하는 추천 시스템으로는 일반적으로 콘텐츠 기반의 추천 시스템과 협업 필터링 기반의 추천 시스템이 널리 사용되고 있다. 콘텐츠 기반의 추천 시스템[18]은 추천 대상 사용자의 이력을 분석하여 아이템의 속성을 찾아내, 이와 유사한 속성을 가진 카테고리의 아이템을 추천해주는 방식이다. 콘텐츠 기반의 추천은 추천 대상자의 이력만 사용하므로, 아이템에 대한 다른 사용자들의 이력을 요구하지 않는다는 장점이 있다. 하지만, 사용자의 과거 이력에 출현한 아이템을 과도하게 추천하는 과대 특수화(Overspecialization) 문제로 인해, 사용자에게 추천되는 아이템의 다양성이 떨어진다는 한계를 갖는다.

협업 필터링[19]은 1992년 처음 소개된 이래 현재도 산업계와 학계에서 여전히 주목받는 추천 방법 중의 하나이다. 협업 필터링은 일반적으로 사용자 기반 또는 아이템 기반 방식으로 구현된다. 사용자 기반이란 추천 대상 사용자와 아이템 구매 이력이 유사한 사용자의 정보를 활용하는 방식이고, 아이템 방식은 추천 대상 아이템에 대해 사용자들의 구매 이력 정보가 유사한 정보를 활용하는 방식이다. 협업 필터링은 사용자에게 다양한 아이템을 추천할 수 있다는 장점을 갖지만 데이터의 희소성(Sparsity) 및 확장성(Scalability) 측면에서 한계를 갖고 있으며, 이러한 한계를 보완하기 위한 연구들이 매우 활발히 진행되고 있다[20-21].

이렇게 다양한 종류의 추천 시스템들이 알리바바(Alibaba), 아마존(Amazon), 구글, 넷플릭스(Netflix), 그리고 유튜브(YouTube) 등 전 세계인들이 사용하는 서비스에 적극적으로 활용되면서, 아이템에 대한 사용자의 리뷰들이 급속도로 증가하는 환경이 조성되고 있다. 또한 트위터(Twitter)나 페이스북(Facebook)과 같은 SNS(Social Networking Service)를 통해 사용자의 이력 정보가 비정형 데이터인 텍스트의 형태로 유통되는 경향이 나타남에 따라, 텍스트 분석을 적용하여 추천 시스템의 성능을 향상시키기 위한 흥미로운 연구들도 수행되고 있다[22].

2. Text Mining and Similarity Analysis

정보기술이 급속도로 발달하고 각 산업에 널리 활용되기 시작하면서, 텍스트 데이터의 규모가 폭발적으로 증가하고 텍스트 분석에 대한 수요 또한 급증하게 되었다. 이에 따라 일반적으로 데이터 마이닝(Data Mining)의 한 분야로 인식되었던 텍스트 분석이 텍스트 마이닝이라는 독립된 연구 분야로 많은 관심을 받게 되었다. 텍스트 마이닝은 뉴스 기사, 블로그, 그리고 SNS 등과 같이 텍스트로 구성된 데이터를 구조화한 후 빈도 분석(Frequency Analysis), 군집화(Clustering), 분류(Classification) 등 전통적인 데이터 마이닝의 주요 개념을 활용하여 분석하는 과정을 일컫는다[23]. <Fig. 3>은 텍스트 마이닝의 과정과 관련 기술을 개괄하여 나타낸다.

<Fig. 3>에서 벡터 공간 모델(Vector Space Model)이란 텍스트의 구조화를 위해 가장 널리 사용되어 온 개념으로, 각 문서를 출현 단어의 가중 빈도(Weighted Frequency) 벡터로 표현한다. 가중 빈도는 문서 내 단어들의 출현 빈도를 그대로 사용하거나 0 또는 1의 이진 값으로 나타낼 수도 있지만, 일반적으로 TF-IDF(Term Frequency - Inverse Document Frequency)를 활용한 가중치 정보를 널리 사용한다[24-25]. TF-IDF란 문서에 출현한 단어 빈도인 TF(Term Frequency)에 해당 단어가 출현한 문서의 빈도에 대한 전체 문서 빈도의 비율인 IDF(Inverse Document Frequency)를 곱하여 산출한다. 즉, 임의의 단어가 자신이 속한 문서에서 출현한 빈도는 높으면서, 전체 문서 중 해당 단어가 출현한 문서의 수는 적을수록 이 단어가 해당 문서에서 갖는 TF-IDF 값은 높게 나타난다.

하지만 일반적으로 분석 대상 문서의 수와 이들 문서에서 출현하는 단어의 수는 매우 많고, 벡터 공간 모델에서 차원의 크기는 단어의 수에 비례하여 증가하기 때문에 벡터 공간 모델을 그대로 사용하여 문서 분석을 수행하기에는 현실적인 어려움이 있다. 따라서 현실적으로 처리 가능한 차원의 크기로 변환하기 위해 PCA(Principal Component Analysis), SVD(Singular Value Decomposition), NMF(Non-negative Matrix Factorization)와 같이 다양한 차원 축소 기법들이 활발하게 적용되고 있다[26-29].

이렇게 구조화된 문서는 분석 목적에 따라 다른 분석 기법과 접목되어 결과 도출에 활용된다. 예를 들면 구조화된 문서에 대한 빈도 분석을 통해 워드 클라우드(Word Cloud)[30], 워드 네트워크(Word Network)[31], 트렌드

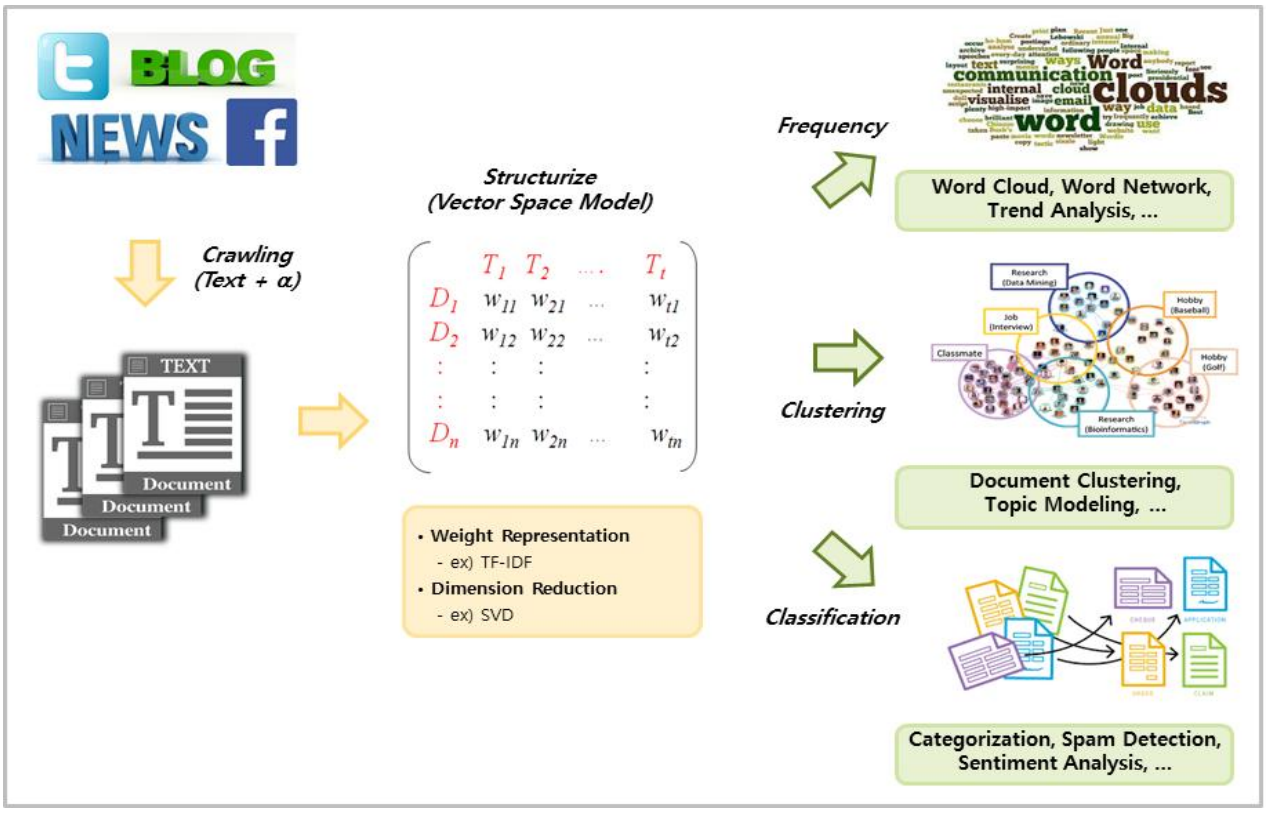


Fig. 3. Text Analytics - Techniques and Applications[23]

(Trend)[32] 분석을 수행할 수 있고, 분류 분석 기법 중 하나인 SVM(Support Vector Machine) 알고리즘[33]을 사용하여 스팸(Spam) 메일을 검출할 수도 있다[34]. 또한 군집 분석을 활용하기 위해 LDA(Latent Dirichlet Allocation)[35] 토픽 모델링을 적용한 연구, K-means 알고리즘[36]을 사용하여 문서를 군집화 한 연구도 수행되었으며, 이외에도, 다양한 기법을 통해 유사 문서를 분석하는 연구들[2-5]이 다수 수행되었다.

일반적으로 문서를 추천하기 위한 방법 중 하나로 문서 간 유사도 정보를 활용하는데, 유사도를 계산하기 위한 대표적인 방법으로 코사인 유사도(Cosine Similarity)가 널리 사용되고 있다. 코사인 유사도는 구조화된 문서의 벡터를 사용하여 두 문서 간 벡터 사이의 코사인 각도를 통해 유사도를 구하는 알고리즘이다. 이외에도 추천을 위해 사용되는 유사도 측정 방식으로 피어슨 상관계수(Pearson Correlation Coefficient), 스피어만 순위 상관계수(Spearman's Rank Correlation Coefficient) 등이 널리 사용되며, 데이터가 희소할 경우 유사 공통 평가 항목만을 비교에 사용하는 자카드 지수(Jaccard Index)도 널리 사용되고 있다[37-39].

본 연구에서는 텍스트 분석을 다룬 다양한 선행 연구의 성과를 활용하여, 사용자의 검색어 정보에 내재된 사용자의 의도를 고려한 유사 문서 추천 방안을 제안한다.

III. Proposed Method

1. Overall Research Process

본 장에서는 사용자 검색어 및 문서 열람 이력 분석을 통해, 사용자의 검색 의도를 반영한 유사 문서를 추천하기 위한 방안을 제시한다. 즉, 사용자 검색 의도를 사용자가 문서를 검색하여 특정 문서를 열람할 때 사용한 검색어로 정의한 후, 사용자의 의도(Intent), 즉 사용자가 문서를 왜(Why) 열람했는지를 고려하여 문서를 추천하는 방안을 새롭게 제안한다. 제안 방법론의 전체 개요는 <Fig. 4>와 같다.

단계 (1)은 사용자별 검색어 및 문서 열람 이력으로부터 실제로 열람된 문서의 유입 검색어를 추출한다. 이 과정을 통해 열람 문서별 유입 검색어 빈도 행렬이 생성되며 자세한 과정은 본 장의 2절에서 소개한다. 단계 (2)는 열람 문서 / 유입 검색어 행렬을 사용하여 열람 문서별 의도 분석을 수행하고, 이를 통해 각 문서별 유입 검색어들에 대해 의도 가중치를 할당하여 열람 문서 / 의도 가중치 행렬을 생성한다. 단계 (3)은 열람 문서 / 의도 가중

치 행렬 정보를 사용하여 문서 간 유사도를 산출한다. 마지막으로 단계 (4)에서는 의도 기반 문서 간 유사도 행렬을 사용하여 사용자가 조회한 특정 문서와 가장 연관 있는 문서를 검색하여 사용자에게 추천한다. 단계 (2) ~ (4)의 과정은 본 장의 3절에서 자세히 소개한다.

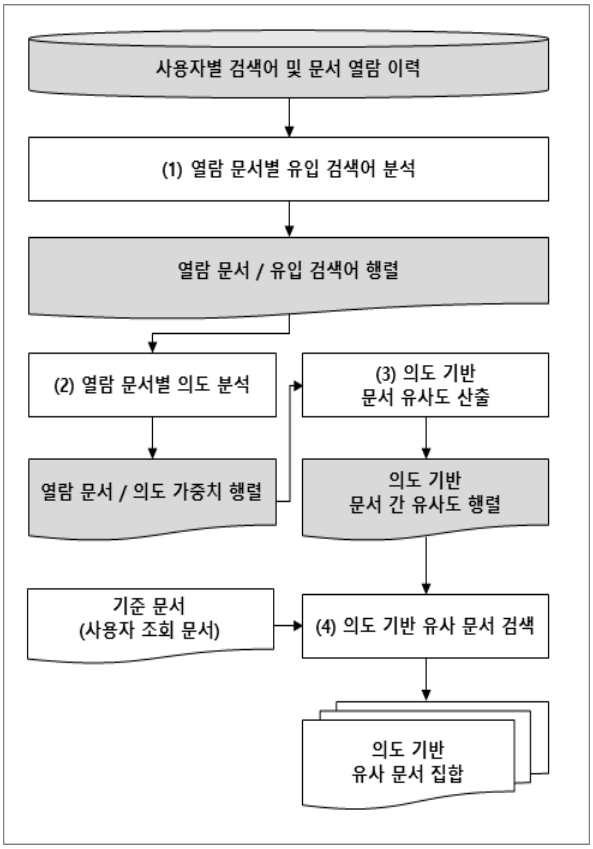


Fig. 4. Overall Research Process

본 장의 이후 절에서는 가상의 예를 통해 본 연구에서 제안하는 방법론을 설명하고, 제안 방법론을 국내 한 전자상거래 플랫폼 기업의 실제 사용자 검색 및 문서 조회 이력 데이터 분석에 적용한 실험 결과는 제4장에서 소개한다.

2. Analyzing Search Keywords for Each Document

본 절에서는 <Fig. 4>의 단계 중 (1)에 해당하는 과정, 즉 사용자별 검색어 및 문서 열람 이력으로부터 열람 문서별 유입 검색어를 분석한 후, 이를 통해 열람 문서 / 유입 검색어 행렬을 생성하는 과정을 소개한다. <Table 1>은 사용자별 검색어 및 문서 열람 이력을 나타낸 가상의 예로, 사용자가 검색을 통해 목록을 획득한 문서 중 실제로 내용까지 열람이 이루어진 문서들의 이력을 나타낸다.

Table 1. History of Search Keywords and Documents Access for Each User

User	Search Keyword	Document
User_1	trade	Doc.1
User_1	polar bear	Doc.4
User_1	ice glacier	Doc.5
User_1	polar bear	Doc.5
User_2	ice glacier	Doc.2
User_2	trade	Doc.2
User_3	bitcoin	Doc.1
User_3	bitcoin	Doc.2
User_4	global warming	Doc.2
User_4	global warming	Doc.3
User_4	ice glacier	Doc.3
User_5	ice glacier	Doc.3
User_5	polar bear	Doc.3
User_5	ice glacier	Doc.4
User_5	trade	Doc.4
User_5	trade	Doc.5

<Table 1>은 User, Search Keyword, 그리고 Document의 세 가지 내용으로 구성된다. User는 검색 후 문서를 열람한 사용자, Search Keyword는 사용자가 입력한 검색어, 그리고 Document는 사용자가 열람한 문서를 의미한다. 검색 결과에는 포함되었지만 사용자가 실제로 클릭을 통해 열람하지 않은 문서는 이력에서 제외된다.

세부적으로 살펴보면 User_1 ~ User_5의 5명의 사용자가 Doc.1 ~ Doc.5의 5개의 문서를 열람하였으며, 이때 “trade”, “polar bear”, “ice glacier”, “bitcoin”, 그리고 “global warming”의 5가지 검색어가 사용되었다. 문서별 열람 빈도는 Doc.1이 2회, Doc.2와 Doc.3이 각각 4회, 그리고 Doc.4와 Doc.5가 각각 3회이다. 이를 문서별 접근 사용자 관점에서 정리한 결과는 <Table 2>와 같고, 동일한 내용을 문서별 검색어 관점에서 정리하면 <Table 3>과 같다.

이렇듯 문서별 열람 빈도를 사용자 관점과 검색어 관점 중 어떤 관점에서 분석하느냐에 따라 문서의 구조화 결과가 상이하게 나타나게 되고, 이는 필연적으로 문서 간 유사도에 영향을 미치게 된다. 이 때, 사용자 관점의

Table 2. Matrix of Documents / Users

Documents	Users					
	User_1	User_2	User_3	User_4	User_5	..
Doc.1	1	0	1	0	0	
Doc.2	0	2	1	1	0	
Doc.3	0	0	0	2	2	
Doc.4	1	0	0	0	2	
Doc.5	2	0	0	0	1	
...						

Table 3. Matrix of Documents / Search Keywords

Documents	Search Keywords					
	trade	bitcoin	global warming	polar bear	ice glacier	
Doc.1	1	1	0	0	0	
Doc.2	1	1	1	0	1	..
Doc.3	0	0	1	1	2	
Doc.4	1	0	0	1	1	
Doc.5	1	0	0	1	1	
...						

구조화는 각 문서들을 누가(Who) 열람했는지에 초점을 두는 반면, 검색어 관점의 구조화는 각 문서들이 어떤 의도로(Why) 열람되었는지에 초점을 두는 것으로 이해할 수 있다. 일반적인 사용자 이력 기반 문서 추천이 <Table 2>의 각 사용자의 문서 접근 이력에 기반을 두어 이루어지는 것과 달리, 본 연구에서는 <Table 3>의 문서별 검색어 기반 유사 문서 추천 방식을 제안한다. <Table 3>의 열람 문서 / 유입 검색어 행렬을 이용하여 이후 분석을 수행하는 과정은 다음 절에서 상세히 소개한다.

3. Analyzing Search Intent for Each Document and Recommending Related Documents

본 절에서는 <Fig. 4>의 단계 중 (2) ~ (4)에 해당하는 과정을 소개한다. 즉 단계 (2)에서 문서 / 유입 검색어 행렬에 TF-IDF를 사용하여 열람 문서 의도를 분석하고, 이를 통해 열람 문서 / 의도 가중치 행렬을 생성한다. 또한 단계 (3)에서 열람 문서 / 의도 가중치 행렬에 코사인 유사도를 사용하여 의도 기반 문서 간 유사도를 산출하고, 이를 통해 의도 기반 문서 간 유사도 행렬을 생성한다. 마지막 단계 (4)에서 사용자가 조회한 기준 문서에 대해 의도 기반 문서 간 유사도 행렬을 참조하여 의도 기반 유사 문서 검색을 수행한다.

단계 (1)에서는 문서별 열람 빈도를 검색어 관점에서 분석하기 위해, 문서별 검색어의 유입 빈도를 계수하여 <Table 3>와 같이 문서를 구조화하였다. 일반적으로 각 문서에 대해 유입 빈도가 높은 검색어가 해당 문서가 갖는 내용을 의미있게 대표한다고 해석할 수 있다. 하지만 고빈도 단어가 항상 해당 문서의 의미를 잘 대표하지는 않는다는 결과가 많은 선행 연구를 통해 알려졌기 때문에, 본 연구에서는 검색어의 단순 유입 빈도가 아닌 TF-IDF 가중 빈도를 구조화에 사용한다.

Table 4. IDF of Each Keyword

Search Keyword	DF	IDF
trade	4건	0.097
bitcoin	2건	0.398
global warming	2건	0.398
polar bear	3건	0.222
ice glacier	4건	0.097
Total DF	5건	

<Table 4>는 Search Keyword, DF, 그리고 IDF의 세 가지 내용으로 구성되며, <Table 3>에 소개한 총 다섯 가지 문서 Doc.1 ~ Doc.5에 검색어가 유입된 빈도 정보를 분석한 검색어별 역문서 빈도(IDF)이다. 키워드별 유입 문서의 수는 “trade”와 “ice glacier”가 각각 4건, “bitcoin”과 “global warming”이 각각 2건, 그리고 “polar bear”가 3건이다. 즉, “trade”, “ice glacier”는 전체 문서 중 80% 이상의 문서에 유입 검색어로 사용되어 가장 낮은 IDF인 0.097이 할당되었으며, 이는 해당 단어가 특정 문서의 열람 의도를 가지는 검색어로 해석되기 어려움을 암시한다. 이와 달리 전체 문서 중 40% 이하의 문서에 검색어로 유입된 “bitcoin”과 “global warming”은 높은 IDF인 0.398이 할당되어, 해당 문서의 열람 의도를 내포하는 검색어로 해석될 수 있다.

<Table 5>는 <Table 4>의 검색어별 IDF 정보를 <Table 3>의 열람 문서별 검색어 유입 빈도인 TF 정보에 곱하여 TF-IDF 가중치를 계산하고, 이를 열람 문서 / 의도 가중치 행렬로 나타낸 것이다. <Table 5>를 통해 Doc.1은 “bitcoin”, Doc.2는 “bitcoin”, “global warming”, Doc.3은 “global warming”, Doc.4와 Doc.5는 “polar bear”가 해당 문서의 열람 의도를 잘 나타내는 검색어임을 알 수 있다.

단계 (3)에서는 <Table 5>에 나타난 열람 문서 / 의도 가중치 행렬을 사용해서 의도 기반 문서 간 유사도 행렬을 생성하며 그 결과는 <Table 6>과 같다. <Table 6>은

Table 5. Matrix of Documents / Intent Weight

Doc.	Search Keywords					
	trade	bitcoin	global warming	polar bear	ice glacier	
Doc.1	0.097	0.398	0.000	0.000	0.000	
Doc.2	0.097	0.398	0.398	0.000	0.097	...
Doc.3	0.000	0.000	0.398	0.222	0.194	
Doc.4	0.097	0.000	0.000	0.222	0.097	
Doc.5	0.097	0.000	0.000	0.222	0.097	
...						

Table 6. Intent-based Documents Similarity

	Doc.1	Doc.2	Doc.3	Doc.4	Doc.5	
Doc.1	1	0.707	0.000	0.088	0.088	
Doc.2	**0.707**	1	0.618	0.124	0.124	...
Doc.3	0.000	0.618	1	0.527	0.527	
Doc.4	0.088	0.124	0.527	1	1.000	
Doc.5	0.088	0.124	0.527	1.000	1	
...						

Table 7. User-based Documents Similarity

	Doc.1	Doc.2	Doc.3	Doc.4	Doc.5	
Doc.1	1	0.231	0.000	0.218	0.436	
Doc.2	**0.231**	1	**0.231**	0.000	0.000	...
Doc.3	0.000	0.231	1	0.436	0.218	
Doc.4	0.218	0.000	0.436	1	0.800	
Doc.5	0.436	0.000	0.218	0.800	1	
...						

기준 문서와 추천 대상 문서를 각각 행과 열로 구성하며, 표에 제시된 값은 두 문서 간 코사인 유사도를 나타낸다. 코사인 유사도는 1 이하의 값을 가질 수 있으며, 기준 문서와 추천 대상 문서가 유사할수록 1에 가까운 코사인 유사도 값을 가진다.

단계 (4) 의도 기반 유사 문서 검색을 설명하기 위해, 사용자가 클릭을 통해 내용을 열람한 기준 문서를 Doc.2로 가정한다. 이때 <Table 6>에서 기준 문서 Doc.2와 가장 높은 유사도를 갖는 추천 대상 문서는 Doc.1로, 0.707의 유사도를 갖는다. 이는 Doc.2를 열람한 사용자들의 검색 의도와 Doc.1을 열람한 사용자들의 의도가 가장 유사하다는 것을 의미하므로, Doc.2를 조회한 사용자에게는 Doc.1을 연관 문서로 추천한다.

<Table 6>의 유사도 결과는 문서별 열람 빈도를 검색어 관점에서 분석한 <Table 3>의 열람 문서 / 유입 검색어 행렬을 기준으로 도출된 결과이다. 한편 검색 의도가 아닌 단순 조회 이력을 기반으로 위의 분석을 수행하는 경우, 즉 <Table 2>의 열람 문서 / 사용자 행렬을 기준으로 문서 간 유사도를 도출하는 경우에는 이와 다른 결과인 <Table 7>의 결과가 도출된다. <Table 7>에서 Doc.2와 가장 유사한 문서는 Doc.1과 Doc.3이므로, Doc.2의 연관 문서로 Doc.1과 Doc.3 두 개의 문서를 동시에 추천하게 된다.

이처럼 본 논문에서 제안하는 의도 기반 문서 추천은 전통적인 사용자 기반 문서 추천과는 상이한 방식으로 이루어짐을 살펴보았으며, 제안 방법론의 우수성은 다음 장의 실험을 통해 소개한다.

IV. Experiment

1. Experimental Model and Environments

본 절에서는 제안 방법론의 우수성을 평가하기 위한 실험 과정 및 결과를 소개한다. 본 실험에서는 국내 최대 전자상거래 플랫폼 기업인 'C' 사의 실제 사용자 검색 이력 중 2020년 1월부터 2020년 11월까지 239,438건의 데이터를 활용하였다.

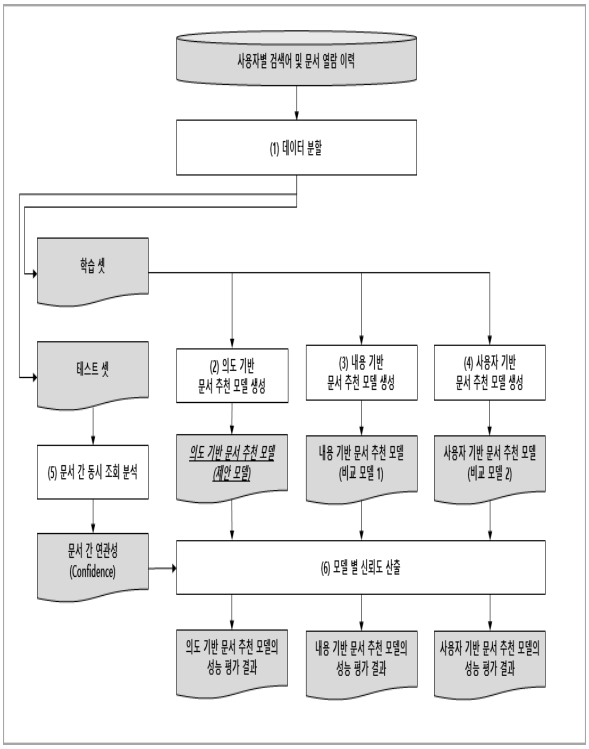


Fig. 5. Experimental Model

실험 모형의 전체 개요는 <Fig. 5>와 같다. 우선 단계 (1)은 사용자별 검색어 및 문서 열람 이력으로부터 학습 셋(Training Set)과 테스트 셋(Test Set)을 분할하고, 단계 (2) ~ (4)는 학습 셋을 사용하여 각각 의도 기반(Intent-based), 내용 기반(Content-based), 그리고 사용자 기반(User-based)의 세 가지 추천 모델을 생성하였다. 다음으로 단계 (5)에서는 각 추천 모델의 성능을 평가하기 위한 기준 생성을 위해 테스트 셋으로부터 문서 간 연관성을 도출하였고, 마지막으로 단계 (6)에서는 단계 (5)에서 도출된 문서 간 연관성 측면에서 각 추천 모델별 성능 평가 결과를 분석하였다.

세부적으로는 사용자별 검색어 및 문서 열람 이력에서 단계 (1)의 데이터 분할을 통해 2020년 1월부터 2020년 8월까지의 이력 167,791건의 학습 셋, 그리고 2020년 9월부터 2020년 11월까지의 이력 71,647건의 테스트 셋

을 구축하였다. 단계 (2) ~ (4)에서는 모델 간 성능의 비교를 위해 의도 기반, 내용 기반, 그리고 사용자 기반의 문서 추천 모델을 생성하였다. 즉, 단계 (2)는 <Table 6>의 의도 기반 문서 간 유사도 행렬을, 그리고 단계 (4)는 <Table 7> 사용자 기반 문서 간 유사도 행렬을 기반으로 문서를 추천한다. 또한 단계 (3)은 문서 간 단순 유사도를 기반으로 문서를 추천한다.

이러한 세 가지 추천 모델은 동일한 문서에 대해서도 각자의 알고리즘에 따라 서로 다른 문서를 연관 문서로 추천하게 된다. 여러 추천 모델의 정확성을 파악하기 위해 본 실험에서는 성능 평가 기준으로 연관분석에서 주로 사용되는 신뢰도를 채택하였으며, 문서 간 신뢰도는 단계 (5)에서 산출하였다. 신뢰도란 특정 사건 A가 발생했을 때 사건 B도 함께 발생할 확률인 조건부 확률로 계산된다. 단계 (5)에서는 단계 (2) ~ (4)에 사용되지 않은 별도의 데이터 셋인 테스트 셋으로부터 문서 간 신뢰도를 계산한다. 본 실험의 단계 (6)에서는 단계 (2) ~ (4)에서 생성된 각 모델별 추천 문서에 대해 단계 (6)의 신뢰도를 산출하여, 의도 기반, 내용 기반, 그리고 사용자 기반 문서 추천 모델의 성능을 평가하였다.

2. Results and Interpretation

본 절에서는 문서 간 연관성을 기준으로 각 추천 모델의 성능을 비교한 실험 결과를 소개한다. <Table 8>은 의도 기반, 내용 기반 그리고 사용자 기반의 추천 모델을 유사도 측정 방식과 문서의 구조화 방식에 따라 각각 4가지 조합으로 세분화하여, 총 12가지의 경우에 대한 실험을 수행한 결과를 요약한 것이다. <Table 8>에서 Rank 1은 각 추천 모델에 의해 특정 문서와 가장 유사한 것으로 식별된 문서 간의 신뢰도를 산출하고, 이러한 과정을 전체 문서에 대해 반복하여 전체 문서와 최유사 문서 간 신뢰도의 평균을 집계한 것이다. 또한 이와 동일한 방법으로 최유사 문서뿐 아니라 유사도 상위 2개 ~ 5개 문서에 대해 실험을 수행한 결과가 Rank 2 ~ 5에 제시되어 있다.

<Fig. 6>은 <Table 8>에 나타난 추천 모델별 Rank 1 ~ 5까지 전체의 평균 신뢰도를 막대그래프로 시각화한 것이다. <Fig. 6(a)>와 <Fig. 6(b)>는 코사인 유사도를 사용하여 유사도를 측정하였으며, <Fig. 6(c)>와 <Fig. 6(d)>는 자카드 지수를 사용하여 유사도를 측정하였다. 한편 <Fig. 6(a)>와 <Fig. 6(c)>는 빈도 기준으로 절대 빈도를 사용하였으며, <Fig. 6(b)>와 <Fig. 6(d)>는 빈도 기준으로 TF-IDF를 사용하였다.

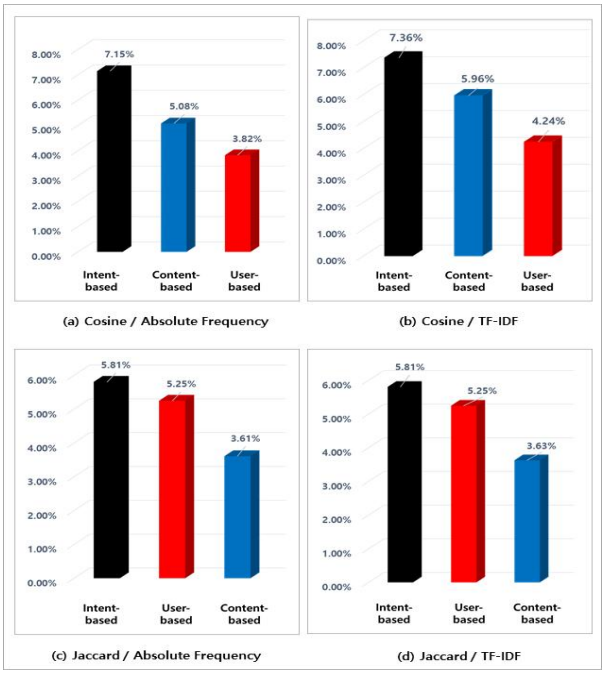


Fig. 6. Overall Average Confidence from Ranking 1 to 5

세부적으로는 본 연구에서 제안하는 의도 기반 추천 모델이 모든 조합에서 가장 우수한 성능을 보였다. 특히

Table 8. Performance Comparison of Three Recommendation Models

Similarity Measure	Frequency	Recommendation Model	Average Confidence of Top N Documents					
			Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	**Avg.**
Cosine Similarity	Absolute Frequency	Intent-based	10.89%	7.44%	6.38%	5.90%	5.13%	**7.15%**
		Content-based	8.44%	5.13%	4.54%	3.90%	3.40%	**5.08%**
		User-based	5.88%	4.41%	3.31%	2.81%	2.71%	**3.82%**
	TF-IDF	Intent-based	10.75%	7.75%	6.94%	6.16%	5.22%	**7.36%**
		Content-based	9.74%	6.66%	5.05%	4.18%	4.17%	**5.96%**
		User-based	6.73%	4.54%	3.61%	3.28%	3.03%	**4.24%**
Jaccard Index	Absolute Frequency	Intent-based	8.06%	6.36%	5.33%	4.95%	4.34%	**5.81%**
		Content-based	6.29%	3.64%	2.95%	2.80%	2.37%	**3.61%**
		User-based	8.84%	5.62%	4.45%	3.98%	3.35%	**5.25%**
	TF-IDF	Intent-based	8.06%	6.36%	5.33%	4.95%	4.34%	**5.81%**
		Content-based	6.28%	3.66%	2.97%	2.79%	2.43%	**3.63%**
		User-based	8.84%	5.62%	4.45%	3.98%	3.35%	**5.25%**

제안 모델은 코사인 유사도와 TF-IDF를 사용한 실험인 <Fig. 6>(b)에서 가장 높은 평균 신뢰도인 7.36%을 나타냈다. 한편 내용 기반 추천 모델과 사용자 기반 추천 모델의 경우 어떤 유사도 기준을 채택하는지에 따라 성능의 순위가 바뀌었으며, 구체적으로는 코사인 유사도를 사용했을 때 내용 기반 추천 모델이, 그리고 자카드 지수를 사용했을 때 사용자 기반 추천 모델이 우수한 성능을 나타냄을 확인하였다. 한편 빈도 기준을 TF-IDF 또는 절대 빈도 중 어떤 것으로 사용하는지는 모델별 성능의 순위에 영향을 주지 않는 것으로 나타났다.

<Table 8>에서 모든 모델은 유사 문서의 범위를 Rank 1에서 Rank 5로 확장할수록 평균 신뢰도가 낮아짐을 알 수 있다. 다만 모델에 따라 평균 신뢰도가 감소하는 정도에는 차이가 있으므로, 이를 비교하기 위해 <Table 8>의 모델별 평균 신뢰도 변화 추이를 <Fig. 7> ~ <Fig. 10>에 도식화하여 나타냈다.

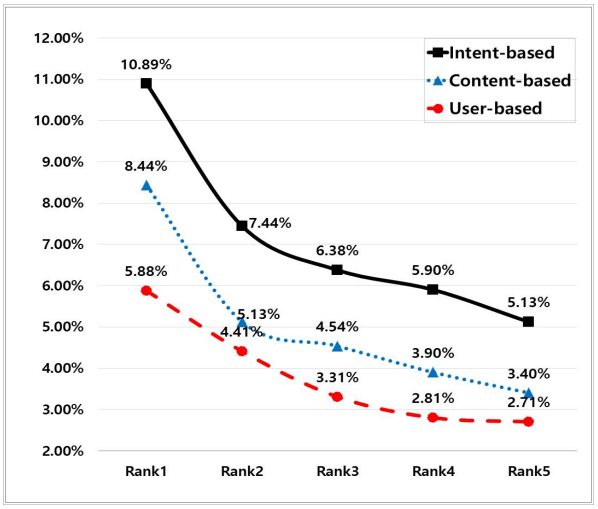


Fig. 7. Average Confidence (Cosine/Absolute Frequency)

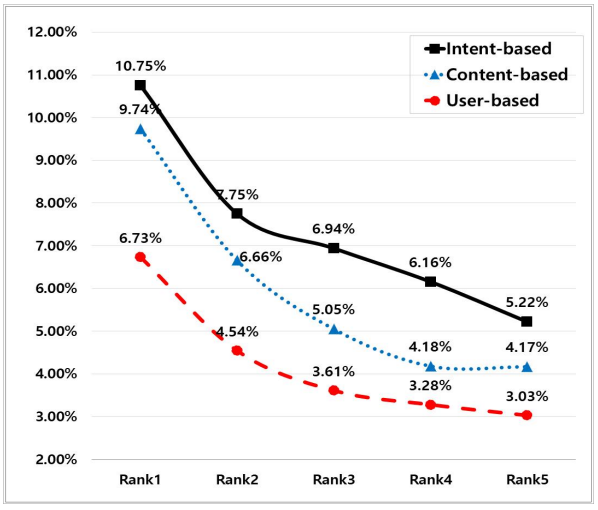


Fig. 8. Average Confidence (Cosine/TF-IDF)

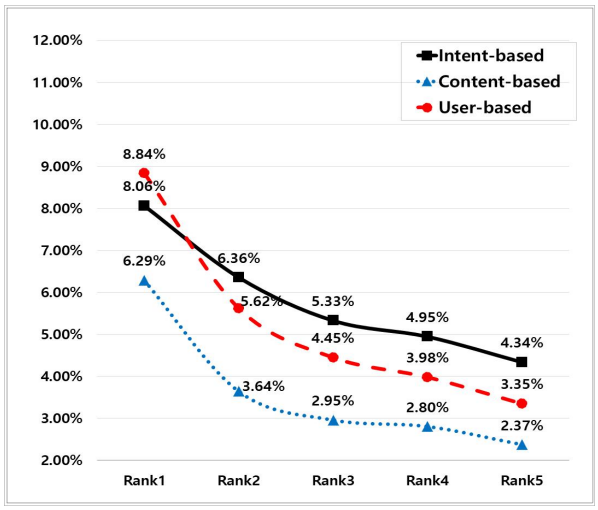


Fig. 9. Average Confidence (Jaccard/Absolute Frequency)

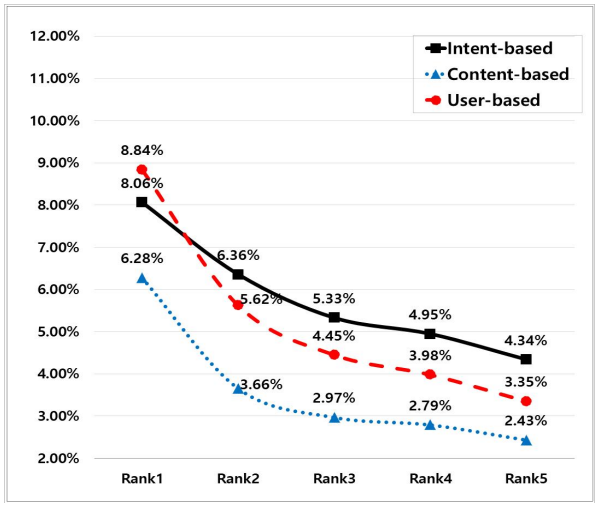


Fig. 10. Average Confidence (Jaccard/TF-IDF)

<Fig. 7> ~ <Fig. 10>에서 최유사 문서, 즉 Rank 1에 대한 평균 신뢰도를 기준으로 평가한 경우 유사도 기준으로 코사인 유사도를 사용했을 때는 제안 방법론인 의도 기반 추천 모델이, 그리고 자카드 지수를 사용했을 때는 사용자 기반 추천 모델이 가장 우수한 성능을 나타냄을 확인하였다. 하지만 자카드 지수를 사용한 경우 사용자 기반 추천 모델은 유사 문서의 범위가 확장됨에 따라 평균 신뢰도가 감소하는 속도가 다른 두 모델에 비해 빠른 것으로 나타났으며, 그 결과 Rank 2 ~ Rank 5의 분석에서는 의도 기반 추천 모델이 오히려 사용자 기반 추천 모델에 비해 우수한 성능을 나타내는 것으로 나타났다. 이렇듯 <Fig. 6>의 전체 평균 신뢰도 비교, 그리고 <Fig. 7> ~ <Fig. 10>의 유사 문서 범위 확장에 따른 평균 신뢰도 비교 결과를 통해, 제안하는 의도 기반 문서 추천 방법론이 기존의 사용자 기반, 또는 내용 기반 추천 모델에 비해 우수한 성능을 나타내는 것을 확인할 수 있었다.

V. Conclusion

사용자들이 방대한 데이터로부터 원하는 정보를 수월하게 획득할 수 있도록 지원하기 위해, 사용자가 접근한 문서와 관련있는 문서를 추천하는 연구들이 다수 수행되었다. 전통적으로 사용자의 문서 열람 이력을 바탕으로 사용자 선호를 고려해 문서를 추천하는 다양한 방법들이 제안되었으나, 이러한 접근법은 문서를 누가 열람했는지의 정보만을 활용할 뿐, 사용자가 해당 문서를 열람하게 된 의도를 충분히 활용하지 못했다는 한계를 갖는다. 따라서 본 연구에서는 사용자가 문서 검색에 사용한 검색어를 활용하여, 사용자의 검색 의도에 기반을 둔 문서 추천 방안을 새롭게 제시하였다. 또한 제안 방법론의 실무적 활용 가능성을 판단하기 위해 국내 전자상거래 플랫폼 기업인 'C' 사의 실제 사용자 검색 이력 239,438건을 분석한 실험을 수행하였으며, 실험 결과 제안 방법론이 기존의 내용 기반 추천 모델 및 단순 열람 이력 기반 추천 모델에 비해 우수한 성능을 보임을 확인하였다.

본 연구의 기여는 다음과 같다. 우선 본 연구는 기존의 내용 기반 혹은 사용자 이력 기반 유사 문서 식별 외에, 문서 열람을 발생시킨 유입 키워드를 활용한 유사 문서 식별 방안을 새롭게 제시했다는 점에서 학술적 기여를 인정받을 수 있다. 즉 검색어는 열람 문서에 비해 사용자의 정보 검색 의도를 더욱 직접적으로 담고 있으므로, 향후 사용자가 입력한 검색어와 사용자가 열람한 문서의 관계를 분석하는 방식의 많은 후속 연구가 이루어질 것으로 기대한다. 또한 실험을 통해 본 연구에서 제안하는 의도 기반 추천 모델이 기존의 내용 기반, 혹은 사용자 이력 기반 추천 모델에 비해 평균 신뢰도 측면에서 우수한 성능을 보임을 확인하였으며, 이는 본 연구의 실무적 기여로 인정받을 수 있다.

본 연구의 한계는 다음과 같다. 우선 본 연구에서 문서 추천을 위해 사용한 검색어 정보는 해당 시스템의 관리 및 운영 권한을 갖지 않은 일반 사용자가 획득하기에는 어려움이 있다. 따라서 제안 방법론을 고도화하고 검증하기 위해서는 실제 시스템을 운영하고 있는 주체의 참여가 수반되어야 하며, 이는 본 연구의 확장성 측면의 한계가 될 수 있다. 또한 본 연구에서는 문서 간 연관성 척도를 사용하여 다양한 문서 추천 모델의 성능을 비교하고 엄밀한 평가를 위해 학습 데이터와 검증 데이터의 기간에 차이를 두었다. 하지만 이러한 평가 방법은 문서 추천 모델의 성능을 직접적으로 평가한 것은 아니라는 한계를 갖는다. 따라서 향후 연구에서는 제안 모델과 비교 모델을 실제 시스템에 적용하여 각 모델의 추천 성능 및 사용자 만족도 향상 정도를 분석할 필요가 있다.

REFERENCES

[1] D. Reinsel, J. Gantz, and J. Rydning, Data Age 2025: The Evolution of Data to Life-Critica, *https://www.import.io/wp-content/uploads/2017/04/Seagate-WP-DataAge2025-March-2017.pdf*

[2] A. Lee, K. Choi, and G. Kim, "LDA Topic Modeling and Recommendation of Similar Patent Document Using Word2vec," Information Systems Review, Vol. 22, No. 1, pp. 17-31, Feb. 2020.

[3] J. Kim, J. Byun, D. Sun, T. Kim, and Y. Kim, "A Model for Measuring the R&D Project Similarity using Patent Information," Journal of the Korea Institute of Information and Communication Engineering, Vol. 18, No. 5, pp. 1013-1021, May. 2014. DOI: 10.6109/JKIICE.2014.18.5.1013

[4] Y. Bai and S. Park, "LEXAI : Legal Document Similarity Analysis Service using Explainable AI," Journal of Computing Science and Engineering, Vol. 47, No. 11, pp. 1061-1070, Nov. 2020. DOI: 10.5626/JOK.2020.47.11.1061

[5] H. Lee and J. Kim, "Issue Keyword Extraction Method Using Document Similarity Method-Focused on Internet Articles -," Asia-pacific Journal of Multimedia services convergent with Art, Humanities, and Sociology, Vol. 7, No. 8, pp. 383-391, Aug. 2017. DOI: 10.35873/ajmahs.2017.7.8.035

[6] J. Kim, J. Suh, D. Ahn, and Y. Cho, "A Personalized Recommendation Methodology based on Collaborative Filtering," Journal of Intelligence and Information Systems, Vol. 8, No. 2, pp. 139-157, Dec. 2002.

[7] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon, and J. Riedi, "GroupLens: applying collaborative filtering to Usenet news," Communications of the ACM, Vol. 40, No. 3, pp. 77-87, Mar. 1997. DOI: 10.1145/245108.245126

[8] S. Lee, Y. Cho, J. Lee, and D. Yu, "Comparative study of recommender systems using movie rating data," Journal of the Korean Data And Information Science Sociaty, Vol. 31, No. 6, pp. 975-991, Nov. 2020. DOI: 10.7465/jkdi.2020.31.6.975

[9] Y. Yoo, J. Kim, B. Sohn, and J. Jung, "Evaluation of Collaborative Filtering Methods for Developing Online Music Contents Recommendation System," The Transactions of The Korean Institute of Electrical Engineers, Vol. 66, No. 7, pp. 1083-1091, Jul. 2017. DOI: 10.5370/KIEE.2017.66.7.1083

[10] T. Shin, K. Chang, and Y. Park, "Customer Recommendation Using Customer Preference Estimation Model and Collaborative Filtering," Korea Intelligent Information Systems Society, Vol. 12, No. 4, pp. 1-14, Dec. 2006.

[11] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," Proceedings of the 1993 ACM SIGMOD international conference on Management of data, pp. 207-216, New York, NY, USA, Jan. 1993. DOI: 10.1145/170035.170072

[12] D. Lee, "A Regression-Model-based Method for Combining Interestingness Measures of Association Rule Mining," Journal of Intelligence and Information Systems, Vol. 23, No. 6, pp.127-141, Mar. 2017. DOI: 10.13088/JIIS.2017.23.1.127

[13] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to Information Retrieval," Cambridge University Press, pp. 1-506, 2008.

[14] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Inormation Retrieval: The concepts and technology behind search (2nd. ed.)," Addison-Wesley Publishing Company, pp. 1-913, 2011.

[15] J. Son, S. Kim, H. Kim, and S. Cho, “Review and Analysis of Recommender Systems,” Journal of Korean Institute of Industrial Engineers, Vol. 41, No. 2, pp. 185-208, Apr. 2015.

[16] C. H. Cai, A. W. C. Fu, C. H. Cheng, and W. W. Kwong, "Mining association rules with weighted items," Proceedings. IDEAS'98. International Database Engineering and Applications Symposium, pp. 68-77, Cardiff, UK, Jul. 1998. DOI: 10.1109/IDEAS.1998.694360

[17] D. Lee, “A Study on the Improvement of Recommendation Accuracy by Using Category Association Rule Mining,” Journal of Intelligence and Information Systems, Vol. 26, No. 2, pp. 27-42, Jun. 2020. DOI: 10.13088/JIIS.2020.26.2.027

[18] Y. Wu and A. Chen, "Index structures of user profiles for efficient web page filtering services," Proceedings 20th IEEE International Conference on Distributed Computing Systems, p. 644-651, Taipei, Taiwan, Apr. 2000. DOI: 10.1109/ICDCS.2000.840981

[19] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," Communications of the ACM, Vol. 35, No. 12, pp. 61-70. Dec. 1992. DOI: 10.1145/138859.138867

[20] Y. Cho and J. Kim, “Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce,” Expert Systems with Applications, Vol. 26, No. 2, pp. 233-246, Feb. 2004. DOI: 10.1016/s0957-4174(03)00138-6

[21] Y. Cho, J. Kim, and S. Kim, “A personalized recommender system based on web usage mining and decision tree induction,” Expert Systems with Applications, Vol. 23, No. 3, pp. 329-342, Oct. 2002. DOI: 10.1016/s0957-4174(02)00052-0

[22] H. Choi and E. Hwang, "Emotion-based Music Recommendation System based on Twitter Document Analysis," KIISE Transactions on Computing Practices, Vol. 18, No. 11, pp. 762-767, Nov. 2012.

[23] N. Kim, D. Lee, H. Choi, and W. X. S. Wong, "Investigations on Techniques and Applications of Text Analytics," The Journal of Korean Institute of Communications and Information Sciences, Vol. 42, No. 2, pp. 471-492, Feb. 2017. DOI: 10.7840/kics.2017.42.2.471

[24] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," Communications of the ACM, Vol. 18, No. 11, pp. 613-620, Nov. 1975. DOI: 10.1145/361219.361220

[25] G. Salton, "The SMART Retrieval System—Experiments in Automatic Document Processing," Prentice Hall, pp. 1-556, 1971.

[26] K. Pearson, "On lines and planes of closest fit to systems of point in space," The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science Series 6, Vol. 2, No. 11, pp. 559-572, Nov. 1901. DOI: 10.1080/14786440109462720

[27] H. Hotelling, "Analysis of a complex of statistical variables into principal components," Journal of Educational Psychology, Vol. 24, No. 6, pp. 417-441, Sep. 1933. DOI: 10.1037/h0071325

[28] G. W. Stewart, "On the early history of the singular value decomposition," SIAM Review, Vol. 35, No. 4, pp. 551-566, Dec. 1993. DOI: 10.1137/1035134

[29] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, Vol. 401, No. 6755, pp. 788-791, Oct. 1999. DOI: 10.1038/44565

[30] O. Park and H. Park, "A Study on the International Research Trends in Electronic Records Management: InterPARES 3 and ITrust Achievements," Journal of Korean Society of Archives and Records Management, Vol. 16, No. 1, pp. 89-120, Feb. 2016. DOI: 10.14404/JKSARM.2016.16.1.089

[31] W. Seo, H. Park, and J. Yoon, "An exploratory study on the korean national R&D trends using co-word analysis," Journal of Information Technology Applications & Management, Vol. 19, No. 4, pp. 1-18, Dec. 2012. DOI: 10.21219/JITAM.2012.19.4.001

[32] H. Choi and H. Varian, "Predicting the present with google trends," Econ. Record, Vol. 88, No. 1, pp. 2-9, Jun. 2012. DOI: 10.1111/j.1475-4932.2012.00809.x

[33] V. Vapnik, "Estimation of Dependences Based on Empirical Data," Springer Verlag, pp. 1-523, 1982.

[34] J. Seo, T. Shon, J. Seo, and J. Moon, "A study on the filtering of spam e-mail using n-Gram indexing and support vector machine," Journal of the Korea Institute of Information Security & Cryptology, Vol. 14, No. 2, pp. 23-33, Apr. 2004.

[35] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," The Journal of Machine Learning Research, Vol. 3, pp. 993-1022, Jan. 2003.

[36] J. Macqueen, "Some methods for classification and analysis of multivariate observations," Proceedings of the Fifth Berkeley

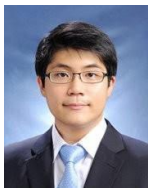
Symposium on Mathematical Statistics and Probability, pp. 281-297, Berkeley, USA, Jun. 1967.

[37] G. Salton and M. J. McGill, "Introduction to modern information retrieval," McGraw-Hill, pp. 1-448, 1983.

[38] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: an open architecture for collaborative filtering of netnews," Proceedings of the 1994 ACM conference on Computer supported cooperative work, pp.175-186, New York, NY, USA, Oct. 1994. DOI: 10.1145/192844.192905

[39] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating Collaborative Filtering Recommender Systems," ACM Transactions on Information Systems, Vol. 22, No. 1, pp. 5-53, Jan. 2004. DOI: 10.1145/963770.963772

Authors



Donghoon Lee received the M.S. degree in Graduate School of Business IT in Kookmin University in 2012. He is currently enrolled in the doctoral program at the same University. He is interested in text mining, deep learning, and data modeling.

Namgyu Kim received the B.S. in Computer Engineering from Seoul National University in 1998, M.S. and Ph.D. degrees in Management Engineering from KAIST, Korea, in 2000 and 2007, respectively. Dr. Kim joined the faculty of the School of Management Information Systems at Kookmin University, Seoul, Korea, in 2007. He is currently a dean of the Graduate School of Business IT at Kookmin University. He is interested in text mining, deep learning, and data modeling.