개인 맞춤형 뉴스 추천 시스템의 설계 및 개발

유영서, 이지민, 이기용
숙명여자대학교 컴퓨터과학부
e-mail : {youngseo, jm94318, kiyonglee}@sookmyung.ac.kr

Design and Development of a Personalized News Recommendation System

YoungSeoYu, Jimin Lee, Ki Yong Lee
Division of Computer Science, Sookmyung Women's University

요 약

실시간으로 뉴스 기사를 제공하는 온라인 뉴스 시스템이 널리 사용되면서, 사람들은 매 순간 속보와 새로운 뉴스 등 대량의 뉴스 기사에 노출되어 있다. 하지만 방대한 뉴스들로부터 사용자가 원하는 뉴스를 찾는 것은 매우 어려운 일이다. 따라서 개인 관심사에 따라 뉴스를 추천해주는 개인 맞춤형 뉴스 추천 시스템의 필요성이 증가되고 있다. 본 논문에서는 사용자의 관심사를 분석하여, 사용자의 관심사에 따라 관련된 뉴스를 자동으로 추천해주는 뉴스 추천 시스템을 설계 및 개발한다. 제안 시스템은 각 사용자가 북마크한 뉴스 기사와 읽은 뉴스 기사를 클러스터링하여 사용자별 프로파일을 생성한다. 또한 전체 뉴스 기사들을 클러스터링하여 주제 별로 분류한다. 사용자에게 뉴스를 추천하기 위해, 제안 시스템은 해당 사용자 프로파일에 포함된 각 클러스터에 대해 전체 뉴스 기사에 대한 클러스터들 중 가장 가까운 클러스터를 찾아 해당 클러스터 내의 뉴스 기사들을 거리 순으로 추천한다. 실제 구현된 시스템을 통해, 제안한 뉴스 추천 시스템이 각 개인에게 뉴스를 효과적으로 추천함을 보인다.

1. 서론

실시간으로 뉴스 기사를 제공하는 온라인 뉴스 시스템이 널리 사용되면서, 사람들은 매 순간 속보와 새로운 뉴스 등 대량의 뉴스 기사에 노출되어 있다. 하지만 방대한 뉴스들로부터 각 사용자들이 원하는 뉴스를 찾는 것은 더욱 어려운 일이 되었다. 따라서 개인 관심사에 따라 개인 맞춤형 뉴스를 추천하는 개인 맞춤형 뉴스 추천 시스템의 필요성이 증가하고 있다. 본 논문에서는 사용자의 관심사를 자동으로 파악하고, 파악된 관심사에 따라 관련된 뉴스를 추천해주는 개인 맞춤형 뉴스 추천 시스템을 설계 및 개발한다.

제안 시스템은 각 사용자가 북마크한 뉴스 기사와 읽은 뉴스 기사를 내용 기반으로 클러스터링하여 사용자별 프로파일을 생성한다. 또한 전체 뉴스 기사들을 내용 기반으로 클러스터링하여 주제별로 분류한다. 제안 시스템은 특정 사용자에게 뉴스를 추천하기 위해, 해당 사용자 프로파일에 포함된 각 클러스터에 대해 전체 뉴스 기사에 대한 클러스터들 중 가장 가까운 클러스터를 찾아 해당 클러스터 내의 뉴스 기사들을 거리 순으로 추천한다. 본 논문은 실제 구현된 시스템을 통해 제안한 뉴스 추천 시스템이 실제 효과적으로 개인 맞춤형 뉴스를 추천함을 보인다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구를 기술한다. 3장에서는 제안하는 개인 맞춤형 뉴스 추천 시스템을 자세히 설명한다. 4장에서는 실제 구현된 제안 시스템의 수행 결과를 보여준다. 마지막으로 5장에서는 결론을 맺는다.

2. 관련연구

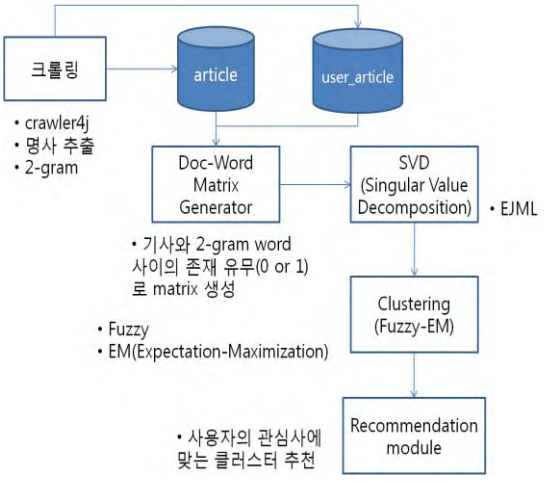
기존의 뉴스 추천 시스템은 주로 협업 필터링 콘텐츠 기반의 방법을 사용한다. 두 가지 방법을 결합하여 사용하는 경우도 있다. 하지만 기존의 방법은 다음과 같은 한계점을 가지고 있다. 협업 필터링은 사용자의 선호도와 비슷한 사용자가 어떤 아이템을 선호하는가를 기반으로 아이템을 추천하는 방법이다. 이러한 협업 필터링의 특성상 다른 사용자가 아직 읽지 않은 뉴스는 추천할 수 없다. 협업 필터링 방법은 충분한 정보를 수집하기 위해서 몇 시간을 기다린 후에야 비로소 사용자에게 추천이 이루어진다[4]. 이것은 특히 뉴스와 같이 신속성이 중요한 분야에서 개선되어야 할 문제점이다.

콘텐츠 기반의 추천 방식은 사용자가 선호하는 아이템과 비슷한 속성을 가진 아이템을 추천하는 방법이다. 콘텐츠 기반 방식은 각 아이템이 가질 수 있는 속성들을 어떻게 정의하느냐에 따라 추천 결과가 매우 달라질 수 있다. 보통 아이템의 속성을 정의하는 과정은 사람이 직접 하는 방법으로 이루어지기 때문에 계속적

인 관리가 필요하다.

Google news의 경우 협업 필터링과 콘텐츠 기반의 방법을 결합하여 사용한다[4]. Google news의 콘텐츠 기반 방법에서는 뉴스의 속성으로 주어진 고정된 카테고리를 사용한다. 카테고리는 국제, 경제, 스포츠 등 9가지로 나뉜다. 카테고리를 이용하여 뉴스를 추천할 경우 두 가지 문제가 있다. 첫째, 세분화 되지 않은 카테고리 분류로 사용자에게 뉴스를 추천할 경우, 추천되는 뉴스가 사용자의 관심사가 아닐 가능성이 크다. 예를 들어, 어떤 사용자가 야구에 관한 기사를 많이 볼 경우, 야구가 속해있는 스포츠 카테고리의 순위가 높아질 것이다. 그렇다면 해당 사용자에게는 스포츠 카테고리에 속하지만 야구와는 관련되지 않은 기사(축구, 배구 등)가 추천될 수 있다. 둘째, 정확하지 않은 카테고리 분류로 추천 결과가 달라질 수 있다. 같은 주제를 다루고 있는 뉴스라 하더라도 서로 다른 카테고리로 분류되는 경우가 존재한다. 이는 뉴스의 주제가 어떤 카테고리에 속해야 하는지 분명하지 않을 때 종종 발생한다. 주요 카테고리로 분류가 되지 않는 뉴스까지도 사용자들에게 추천이 될 수 있어야 한다.

이러한 문제점을 해결하기 위해 본 논문은 사람이 직접 카테고리를 정의하지 않아도, 뉴스를 각 사용자의 관심사에 따라 내용 기반으로 추천하는 개인 맞춤형 뉴스 추천 시스템을 제안한다.



<그림 1> 시스템 구조

3. 제안 뉴스 추천 시스템

본 논문에서 제안하는 개인 맞춤형 뉴스추천 시스템의 구조는 <그림 1>과 같다. 다음은 각 단계를 상세히 설명한다.

① Crawling 단계

제안 시스템은 우선 웹을 crawling하여 뉴스 기사들을 수집한다. 본 논문에서는 포털 사이트인 다음(Daum) 뉴스(media.daum.net)로부터 뉴스 기사를 수집하였으며, crawling을 위해서 Java 기반의 Crawler4j 라이브러리[2]를 사용하였다.

② 단어 추출 단계

뉴스 기사를 crawling한 뒤에는 각 뉴스 기사로부터 URL, 기사 제목, 기사 내용을 추출해 낸다. 그리고 이로부터 다음과 같은 3단계를 통해 각 기사를 대표하는 단어(word)들을 추출한다. 첫 번째 단계에서는 형태소 분석을 통해 기사 제목 및 기사 내용에서 명사를 추출한다. 본 논문에서는 한나눔 라이브러리를 통해 명사를 추출하였다[5]. 명사 중에서도 보통명사, 고유명사, 외국어만을 추출했다. 의존명사, 대명사(예: 너, 우리, 거기, 무엇), 수사(예: 첫째, 하나)는 보편적으로 많이 등장하는 단어이고 기사의 내용을 대표하는 단어가 되기 어렵기 때문에 제외하였다. 두 번째 단계에서는 앞서 추출한 명사들의 tf-idf 가중치를 계산한 뒤 계산된 가중치가 높은 상위 명사를 선택한다. 세 번째 단계에서는 추출한 명사들을 2-gram으로 분할한다. 이것은 뉴스 간의 유사도를 계산할 때 복합명사의 띄어쓰기 문제를 해결하기 위한 것이다. 여기서 2-gram을 선택한 것은 1-gram, 3-gram 기반으로 단어를 분할하는 것보다 뉴스 간 유사도를 파악하는 데 효율적이기 때문이다. 1-gram의 경우 '분산교환망'과 같은 단어에서 '산'이나 '분'과 같은 부적합한 단어들이 과도하게 추출되어 효과가 떨어진다. 3-gram의 경우에는 '정보 검색'과 '정보검색'과 같은 단어의 경우 각각 {정보검, 보검색}과 {정보, 검색}으로 분할되어 둘은 공통된 단어가 추출되지 않는다. 따라서 2-gram 분할을 선택하였다[4]. 이러한 과정을 거치면 각 뉴스 기사를 대표하는 단어들이 <그림 2>의 예와 같이 2-gram 형태로 추출된다.

뉴스 ID	단어
1	지카, 카바, 바이, 이러, 러스, …
2	한국, 국인, 인첫, 첫지, 지카, …
3	환자, 자가, 가발, 발생, 생함, …
4	바이, 이러, 러스, 스에, 감염, …

<그림 2> 2-gram 단어 추출 예

③ Document-word matrix 생성 단계

앞 단계의 결과로 얻은 각 뉴스 기사의 대표 단어들을 document-word matrix로 표현한다. <

그림 3>의 예와 같이 각 뉴스 기사에 해당 단어가 존재하면 1, 존재하지 않으면 0으로 나타낸다.

	지카	카바	바이	이러	러스	한국	…
1	1	1	1	1	1	0	…
2	1	0	0	0	0	1	…
3	0	0	0	0	0	1	
4	0	0	1	1	1	0	

<그림 3> Document-word matrix 생성 예

④ Singular Value Decomposition (SVD) 단계

SVD로 document-word matrix를 분해하여 각 뉴스 기사를 그의 특징 성분들로 구성된 벡터로 나타낸다. SVD를 수행하면 document-word matrix M은 M = UΣV형태로 분해되며, 여기서 U의 각 행은 각 뉴스 기사를 그의 성분으로 분해한 벡터를 나타낸다. 본 논문에서는 계산의 효율을 위해 U의 열 중 중요도가 높은 순으로 5개의 열을 취하여 5차원의 특징 벡터를 생성하였다. 이에 따라 각 뉴스 기사는 최종적으로 5차원의 벡터로 표현된다.

⑤ 클러스터링 단계

이 단계에서는 앞서 5차원 벡터로 표현된 모든 뉴스 기사들을 모아 클러스터링을 수행하여 주제 별로 나눈다. 본 논문에서는 클러스터링 알고리즘으로 Fuzzy-EM[1]을 사용한다. 클러스터링 알고리즘에는 *k*-means라는 많이 사용되는 클러스터링 알고리즘이 있다. 하지만 *k*-means의 경우 클러스터링을 하기 위하여 미리 클러스터의 개수 정해주어야 하는 문제가 있다. 따라서 뉴스기사의 개수에 따라 적적한 클러스터의 개수를 정해주어야 하는 문제가 생긴다. 그러므로 본 논문에서는 미리 클러스터의 개수를 정해줄 필요 없이 입력 데이터에 따라 클러스터의 개수를 자동적으로 정해주는 Fuzzy-EM를 사용하였다. Fuzzy-EM은 먼저 뉴스 기사를 나타내는 각 벡터들 중 임의의 점들을 선택하여 각 클러스터의 중심점으로 한다. 그리고 다음의 과정을 중심점의 변화가 미세해질 때까지 반복한다. (1) E-step에서는 모든 점에 대해 각 클러스터에 대한 소속 확률을 계산한다. (2) M-step에서는 표준 편차를 최소화 할 수 있도록 클러스터 중심을 재계산한다.

⑥ 사용자 프로파일링 단계

본 논문에서는 사용자의 관심사를 파악하기 위해 사용자가 북마크한 뉴스 기사와 사용자가 클릭하여 읽은 뉴스 기사를 사용한다. 이를 위해 제안 시스템은 사용자가 북마크한 뉴스 기사와 사용자가 읽은 뉴스 기사에 대해 앞서 설명한 ②~⑤단계와 동일한 과정을 거쳐 해당 기사들에 대한 클러스터링 결과를 얻는다. 이 결과에 포함된 클러스터들은 각각 해당 사용자의 관심사들을 나타낸다.

⑦ 개인 맞춤형 뉴스 추천 단계

앞서 설명한 단계를 통해 전체 뉴스 기사에 대해 생성된 클러스터들의 집합을 C라고 하고, 특정 사용자에 대해 생성된 클러스터들의 집합을 U라고 하자. 제안 시스템은 해당 사용자에게 다음과 같은 방법으로 뉴스 기사를 추천한다.

(1) U에 속한 클러스터들 중 가장 큰 클러스터부터 작은 클러스터 순으로 다음과 같이 추천을 진행한다.
(2) U의 각 클러스터 u에 대해, u와 가장 가까운 C의 클러스터 c를 찾는다. 여기서 클러스터 간의 거리는 두 클러스터의 centroid 간의 거리로 정의한다.
(3) c의 속한 각 뉴스 기사들을, u의 centroid와의 거리가 가까운 순서부터 추천한다.

위 방법을 사용하면, 해당 사용자의 관심사와 관련이 있는 뉴스 기사부터 추천이 이루어지게 된다.

4. 실제 수행 예

본 장에서는 제안하는 개인 맞춤형 뉴스 추천 시스템을 구현하고 그 실제 수행 결과를 보인다. 본 논문에서는 제안하는 뉴스 추천 시스템을 구현하기 위해서 Linux 환경에서 Java 및 PHP를 사용하였다. 각 사용자는 안드로이드용으로 개발된 안드로이드 앱을 통해 구현된 뉴스 추천 시스템 서버에 접속하여 뉴스를 추천받을 수 있다. 사용자가 안드로이드 폰에서 웹 브라우저를 통해 북마크를 하거나 안드로이드 앱에서 뉴스를 클릭할 때마다 서버에서는 사용자의 프로파일을 갱신한다.

<그림 4>는 사용자가 북마크한 뉴스 기사들을 보여주는 화면이다. 사용자가 북마크한 기사에는 '더불어민주당과 공천'에 대한 뉴스 5개와 '야구'에 대한 뉴스 3개가 포함되어 있다. 이들을 3장에서 설명한 방법에 따라 클러스터링하면 '더불어민주당과 공천'에 대한 클러스터(c_1)와 '야구'에 대한 클러스터(c_2)가 생성된다. c_1의 크기가 c_2보다 크기 때문에 제안 시스템은 c_1와 가까운 뉴스를 먼저 추천하고, 그 다음 c_2와 가까운 뉴스를 추천한다.

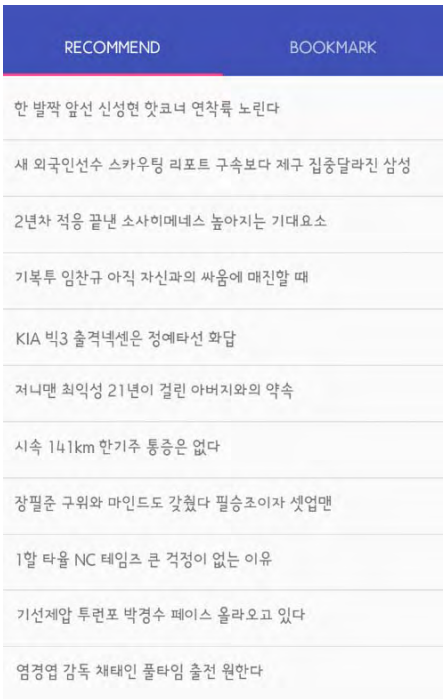
<그림 5>는 c_1과 관련되어 추천된 상위 뉴스 중 일부를 보여준다. 결과를 보면 세 번째와

다섯 번째 뉴스를 제외하고는 더불어민주당 혹은 관련 선거 뉴스인 것을 확인할 수 있다. <그림 6>은 c_2와 관련되어 추천된 상위 뉴스 중 일부를 보여준다. 모두 야구에 관한 뉴스임을 알 수 있다.

실제 구현된 시스템의 수행 결과를 통해, 사용자가 '더불어민주당과 공천', '야구'에 대한 뉴스를 북마크하거나 읽는 경우, 해당 주제에 관한 뉴스가 추천됨을 확인할 수 있었다.

<그림 4> 사용자가 북마크한 뉴스의 예

<그림 5> 첫 번째 클러스터에 대한 추천 뉴스



<그림 6> 두 번째 클러스터에 대한 추천 뉴스

5. 결론

본 논문에서는 개인 맞춤형 뉴스 추천 시스템을 설계하고 실제 구현된 결과를 보였다. 유사한 사용자나 정해진 카테고리를 이용하는 기존의 뉴스 추천 시스템과 달리, 제안 시스템은 뉴스의 내용을 분석하여 사용자가 관심을 가지고 있는 뉴스의 주제와 비슷한 뉴스를 추천해 준다. 또한 사용자의 관심 뉴스가 추가될 때마다 이를 실시간으로 사용자 프로파일에 반영함으로써 추천의 실시간성을 높일 수 있다.

Acknowledgement

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2015R1C1A1A02037071)

참고문헌

[1] Jiawei Han, Micheline Kamber, Jian Pei, "DATA MINING Concepts and Techniques", Elsevier Science Ltd, 2011
[2] https://github.com/yasserg/crawler4j
[3] 이준호, 안정수, 박현주, 김명호, 한글 문서의 효과적인 검색을 위한 n-Gram 기반의 색인 방법, 정보관리학회지, vol. 13, no.1, pp.47-63, 1996.
[4] J. Liu et al, Personalized News Recommendation Based on Click Behavior, In Proceedings of the 15th Int'l Conf. on Intelligent User Interface, pp. 31-40, 2010.
[5] http://kldp.net/projects/hannanum