추천시스템의 이해와 문서 유사도

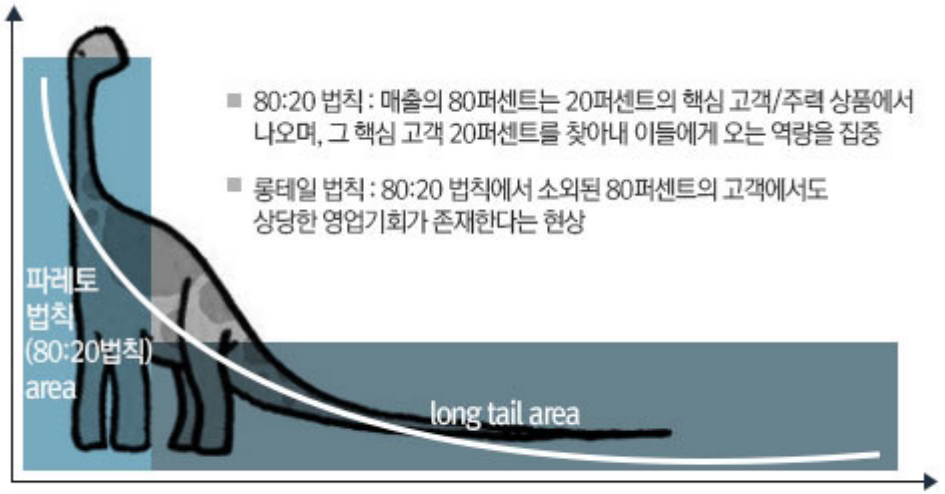
추천시스템이란

개인 맞춤형 서비스 제공 위해 구매패턴 등 과거 데이터를 분석하여 상품을 추천하는 시스템

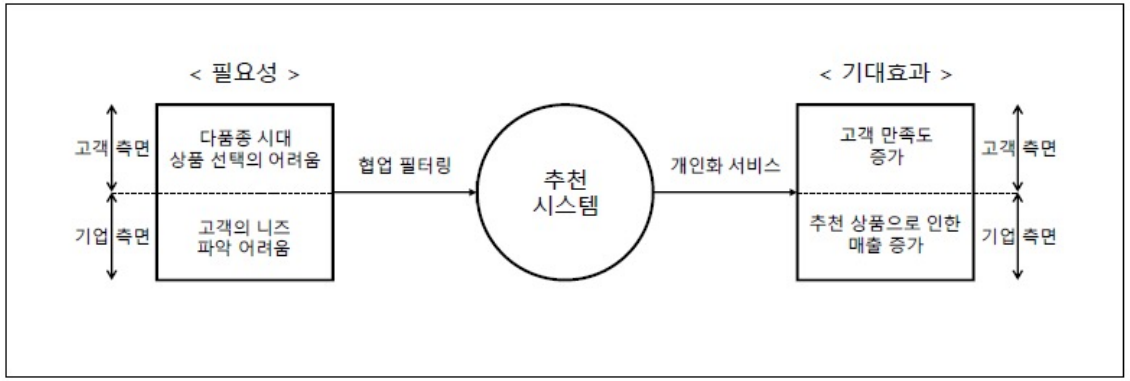
추천시스템은 왜 뜨고 있을까?

고객 선택 및 구매 적중률을 높이기 위해 데이터에 대한 메타 정보 관리와 분석 알고리즘이 중요해지고 있다.

롱테일 문제를 해결할 수 있다



고객, 기업 양측 모두에게 이익이 된다



추천시스템은 여러 가지 종류가 존재한다

여러가지 방식의 추천시스템들이 존재한다.

1. 사용자 프로파일링 기반

2. Segment 기반

3. 상품 연관규칙 기반

4. CF(협업 필터링) 기반

5. CBF(컨텐츠 베이스 필터링) 기반

6. 딥러닝 기반

개인화 콘텐츠 추천 알고리즘 유형

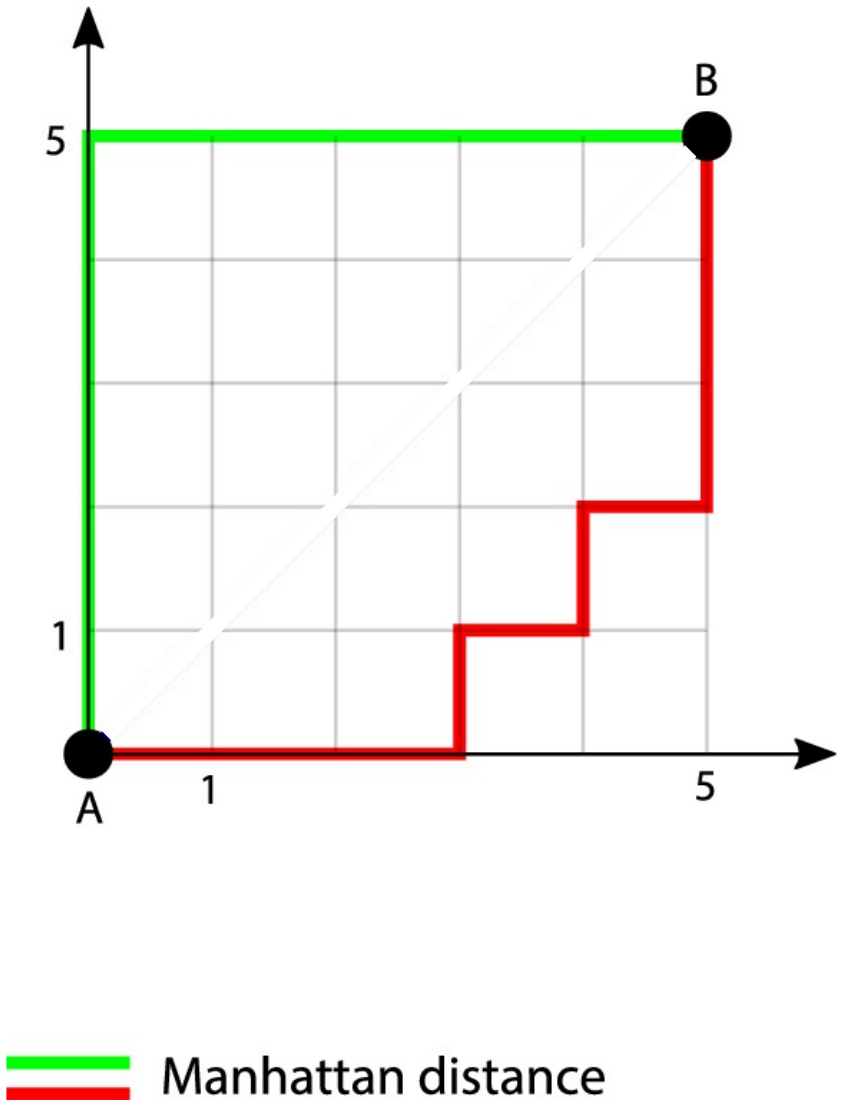
구분	알고리즘	알고리즘 상세 설명	한계		한계 극복 방안	
전통적 알고리즘	CF(협업 필터링)	– 사용자 행동 분석 – 아이템기반, 사용자기반 CF - 행렬분해(MF) 잠재요인 CF	콜드 스타트	– 초기 정보 부족의 문제점 – 새로운 항목 추천 한계	- CBF - 딥러닝 기반 필터링	– 항목 자체 내용 분석 기반 – KNN, DBSCAN 등 AI기술
			계산 효율 저하	– 다수 사용자의 경우 비효율 – 행렬 분해 시 장기간 계산	– 병렬 컴퓨팅	– 행렬 계산 최적화 컴퓨팅 사용 – GPGPU, Grid Computing 등
			롱테일 문제	– 비대칭적 쏠림현상 발생 – 관심 저조 항목 정보 부족	- CBF	– 자료 내 사용자 패턴기반 추천 – LDA, 베이지안 네트워크
	CBF(콘텐츠기반 필터링)	– 콘텐츠 내용 분석 – 유클리디언 거리, 코사인 유사도 측정	메타 정보 함축 한계	– 한정된 메타정보로 사용자와 상품의 프로파일 함축 불가	– CF	– 서로 다른 분야 수치 계산 – 피어슨, 자카드 유사도 측정
최신 알고리즘	딥러닝 기반 필터링	– 구글 Text 자동 생성 기술 – 지도/비지도학습 기반 알고리즘	블랙박스	– 딥러닝이 가진 태생적 한계로 내부 알고리즘 해석이 어렵다.	– 설명 가능한 AI - 컴퓨팅 파워 증대	

문서 유사도
거리 계산 방법

Manhattan Distance (L1 distance)

$$\mathrm{d_{L1}}(w, v) = \sum_{i=1}^{d} |w_i - v_i|$$

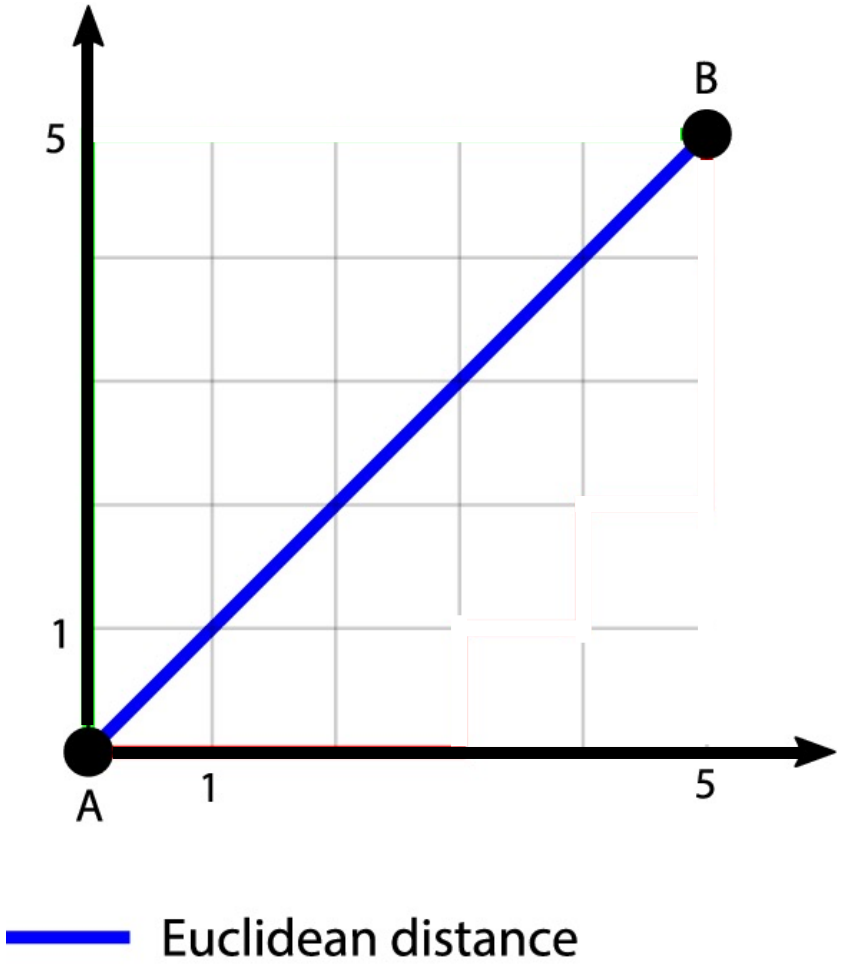
$$\text{where } w, v \in \mathbb{R}^d$$



Euclidean Distance (L2 distance)

$$\mathrm{d}_{\mathrm{L2}}(w, v) = \sqrt{\sum_{i=1}^{d} (w_i - v_i)^2}$$

$$\text{where } w, v \in \mathbb{R}^d$$



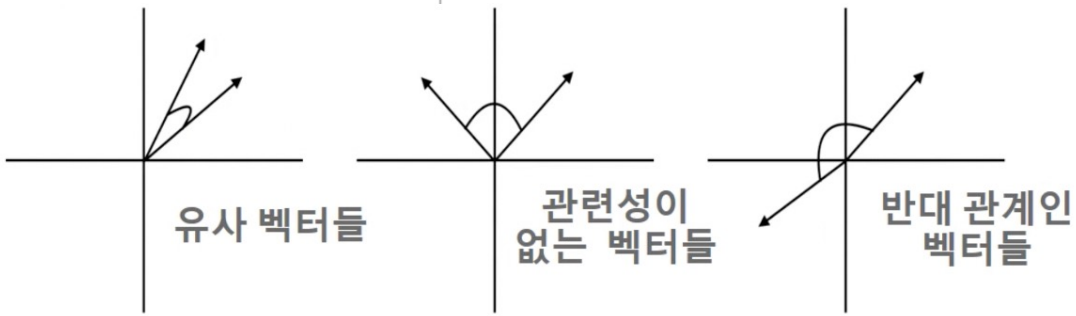
Cosine Similarity

Cosine Similarity는 벡터의 방향을 중요시 함
- Feature vector의 각 차원의 상대적인 크기가 중요할 때 사용

$$\text{similarity} = cos\theta = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

코사인 유사도는 벡터 크기의 비교가 아닌 벡터 방향성의 비교에 중점

	머신러닝	부스팅	책팝니다
D0	50	5	
D1	45		10
D2	10	2	



0도 ~ 90도 Cosine 90도 = 0 Cosine(180) = -1