추천 시스템:
구축, 배포 및 최적화
모범 사례

NVIDIA

요약

본 보고서는 소매, 미디어, 전자 상거래 등 추천 시스템 (recsys) 을 활용하는 유명 기업들의 기술 리더들을 대상으로 실시한 인터뷰 내용을 기반으로 작성되었습니다. 본 연구에 참가한 업체는 Tencent부터 The New York Times까지 다양하게 있습니다.

본 보고서의 대상 독자는 실제 여러 활용 사례에 대한 추천 시스템을 구축 중이거나 구축을 고려하고 있는 데이터 사이언티스트 및 머신 러닝 엔지니어 등입니다. 주된 목표는 해당 분야 전문가의 유용한 인사이트를 제공하고, 추천 시스템 구축, 배포 및 최적화를 위한 모범 사례를 구체화함으로써 보다 광범위한 업계의 논의에 기여하는 것입니다.

> 본 연구를 통해 확인된 주요 가정은 다음과 같습니다.

1. 연관성 있는 추천 시스템을 구축하는 프로세스는 어렵습니다.
2. 어떤 방법이 가장 잘 효율적인지 공유하는 업계 내부의 개방성은 해당 업계의 발전에 필수적입니다.

> 따라서 보고서는 다음 내용을 보여주도록 구성되었습니다.

1. 업계의 추천 시스템 활용 동향 미리보기
2. 추천 시스템이 1979년부터 2009년 사이에 학술적인 연구에서 시작되어 대규모 상업적 성공으로 발전하기까지에 대한 간략한 역사
3. 전문가 인터뷰 발췌문
4. 지금까지의 관찰 동향과 향후 나아갈 방향에 대한 시사점
5. 추천 시스템 관행의 맥락과 미세한 차이를 탐구하는 심층 인터뷰

추천 시스템 동향 미리보기

추천 시스템 분야에서 오픈 소스는 더 이상 선택 사항이 아니라는 점은 이미 명백한 사실입니다. 오픈 소스는 이제 필연적인 것이 되었습니다. 오픈 소스 에코시스템 내에서 도구의 상호 운용성은 특히 전체 수명 주기를 고려할 때 프로젝트의 위험을 낮추는 데 필수적입니다. 이러한 도구는 반드시 실험과 탐구를 뒷받침할 정도로 유연하면서 아직 주목받지 않은 신기술도 받아들일 수 있어야 합니다.

또 한 가지 고려할 사항은 데이터 문제입니다. 2000년대 중반에 '빅 데이터' 개념이 등장한 이래, 데이터의 품질이 뛰어나고 양이 많을수록 더 좋은 모델이 나온다는 것은 머신 러닝에서 보편적으로 인정되고 있습니다. 물론 이는 추천 시스템에서도 마찬가지입니다. 모델을 만들 때 더 우수한 알고리즘에만 의존할 수는 없습니다. 좋은 훈련 데이터, 추론을 위한 좋은 고객 데이터, 좋은 피드백, 메트릭 평가를 위한 도구 등은 반드시 필요합니다. 효과적인 데이터 준비를 위한 우선 순위 또한 변하지 않는 요소입니다. 또한 추천 시스템 활용 성공 사례가 많아질수록 데이터 품질 또한 증가할 것입니다. 메트릭이 성공에 맞춰 개선될수록 데이터로부터 학습하는 것이 중요해집니다. 개발 중인 추천 시스템을 철저하게 이해할 수 있도록 데이터 사이언스 직원을 '메트릭 엔지니어'로 재편성하는 것도 도움이 될 수 있습니다.

> 이 두 가지 외에 본 연구에서 밝혀진 중요한 내용들은 다음과 같습니다.

1. 추천 시스템을 막 도입하려는 기업에 대한 전문가의 조언은 거의 모두 일치했습니다. 유행하는 복잡한 머신 러닝 모델을 구현하려고 서두르지 말고 우선 간단한 것부터 시작해서 무엇이 필요한지 고민하라는 것입니다.
2. 실제 프로덕션에 사용되는 메트릭과 목적 함수를 면밀하게 파악하는 것이 중요합니다
3. 지금까지 추천 시스템은 포인트 솔루션으로 여겨졌지만, 점차 핵심 비즈니스 운영으로 확장되고 있습니다. 기존에 마진이 낮았던 업종에서 고객 신뢰도와 충성도를 높이면서 비즈니스를 성장시키기 위해 추천 시스템을 사용하고 있습니다.
4. 많은 추천 시스템 활용 사례를 보면 보통 100밀리초 미만의 빠른 추론을 필요로 합니다. 이러한 제약 조건 하에서 계속 발전하며 보급되는 AI 기술들이 충분히 빠르게 실행될 수 있을까요? 하드웨어 가속이 이를 가능케 하는 요소가 되었습니다.
5. 업계에서 추천 시스템 관행과 관련하여 주목받는 주제로는 피쳐 스토어 (feature store), 그래프 신경망, 그리고 모델 증류 활용 등이 있습니다.
6. 교훈: 시스템 업그레이드 및 증가하는 활용 사례의 로드맵을 준비하세요. 가능한 경우, 현재의 확장 요구사항뿐만 아니라 성공적인 비즈니스를 위해 몇 년 간의 발전을 위한 방법을 제공하는 기술을 활용합니다.

추천 시스템: 기원 알아보기

업계의 추천 시스템 활용 사례을 살펴보면 당면 과제, 우려 사항, 향후 연구할 영역 등 공통적인 측면이 많습니다. 마찬가지로 추천 시스템의 역사를 살펴보면, 새로운 팀들이 각각 시작한 여정으로부터 배워감에 따라 수십 년에 걸쳐 반복되어 온 테마를 확인할 수 있습니다.

이 섹션에서는 추천 시스템 분야를 확립하는 데 기여한 주목할 만한 프로젝트를 확인할 수 있습니다. 특히 초기의 학술적 실험부터 대규모 상업적 성공으로 이어지기까지의 발전 과정을 추적합니다. 과거의 패턴으로부터 배우고, 이러한 역사를 바탕으로 현재 활용에의 주요 사안과 테마를 비교할 수 있습니다.

1990년대 이전: Grundy, 딥 모델, 책

온라인 추천 시스템의 개념은 1990년대에 등장하기 시작했으며, 이는 전세계 웹의 전반의 성장과 밀접하게 연관되어 있습니다. 초기 프로젝트 중에 Elaine Rich가 개발한 Grundy는 Rich 박사가 1979년 Carnegie Mellon 대학에서 박사 학위 연구 중 창안하고 University of Texas at Austin에서 교수직을 맡아서도 계속한 연구입니다. Rich 박사는 고정관념을 기반으로 사용자 모델을 구축하고 일반화하는 방법을 탐구했습니다. 이 시스템은 사용자에게 (사서로서) 소설을 추천하며, 그 후 사용자는 그 추천의 품질에 대한 피드백을 제공합니다. 이 연구는 비교적 머신 러닝이 초기 단계일 때 이뤄졌지만 특징 엔지니어링, 데이터 훈련, 모델 평가, 모델의 반복 개선을 위한 피드백 등 핵심 요소를 갖추고 있었습니다. 1970년대 후반 AI 학계의 지배적인 견해를 고려하여 Grundy 시스템은 각 사용자에 대한 비교적 깊은 정보에 의존하면서 시스템 사용자 참여를 최소화하려고 시도했습니다.

1990년대: 협업 필터링의 부상

Doug Terry는 1992년 Xerox PARC에서 Tapestry라는 최초의 협업 필터링을 개발했습니다. 이 프로젝트는 이메일 과부하가 발생할 때 발생하는 문제점을 해결했습니다. 사용자가 이메일 메시지에 주석을 달아 중요도를 표시하면 시스템이 주석을 통해 사용자의 중요도를 학습합니다. Terry는 Tapestry에 대한 원본 논문에서 Tapestry의 주요 기술 혁신은 예측 가능한 시멘틱 (semantic) 을 가진 필터 쿼리에 대한 효율적인 알고리즘이라고 설명했습니다. 1992년에 이미 Terry는 앞으로 협업 필터링에서 중요해질 핵심 요소를 파악했습니다. 많은 훈련 데이터 필요성, 보안 및 개인 정보 보호 문제, Tapestry가 새로운 '웹 브라우저' 소프트웨어에 통합되지 않았다는 사실 등입니다.

2년 뒤인 1994년 Paul Resnik은 동료들과 함께 MIT (Massachusetts Institute of Technology) 에서 GroupLens 프로젝트를 시작했으며, 이는 현재 University of Minnesota에서 계속되고 있습니다. 이들의 연구는 Terry의 Tapestry 프로젝트를 개량하여 온라인 뉴스 콘텐츠에 대한 협업 필터링을 제공하는 것이었습니다. 사람들이 온라인 뉴스 기사를 읽고 반응할 때의 데이터가 대체로 무시된다는 점에 착안한 연구였습니다. Elaine Rich가 Grundy에대한 관점과 달리 GroupLens는 비교적 얕은 데이터를 대량으로 활용했습니다. 또한 오픈 프로토콜을 사용했기 때문에 누구나 뉴스 클라이언트를 수정하여 등급과 예상 점수를 통합할 수 있었습니다.

MIT, Firefly Music 추천, 전자 상거래

추천 시스템에 대한 관심은 MIT에서 확산됐습니다. MIT의 Pattie Maes 교수와 대학원생 Upendra Shardanand는 음악 추천 시스템 Ringo를 개발했습니다. 특히, 이 시스템은 보다 개인 맞춤형 추천을 제공하여 사용자가 2000명 이상으로 확장되었습니다. Ringo 개발자들은 Rich의 사용자 모델링용 고정관념을 참고하고, 벤치마크를 통해 자신들의 핵심 알고리즘이 Resnik과 회사의 GroupLens보다 우수하다는 것을 보여주고자 했습니다. 그들은 또한 권장 결과 속도를 높이고 콘텐츠 모델링을 실험하는 데 도움이 되도록 데이터에 클러스터링 알고리즘을 사용하는 등 향후 머신러닝의 필요성에 주목했습니다.

후속 프로젝트에서 Pattie Maes와 동료들은 1995년 MIT에서 Firefly Music 추천 시스템을 출시했습니다. 이는 벤처 지원 스타트업으로 분리되었고, 1998년 Microsoft에 인수되었습니다. 이 회사는 음악 추천뿐 아니라 책, 뉴스 등 다른 콘텐츠로도 사업을 확장했으며, 자신들의 기술 라이선스를 Yahoo!, ZDNet, Barnes & Noble, America Online, Reuters 등 인터넷 시대 초기의 여러 영향력 있는 기업들에 제공했습니다.

Firefly는 머신 러닝을 전자 상거래에 활용한 최초의 추천 시스템으로 알려져 있습니다. 상업적 성공을 거둔 초기 추천 시스템 연구 프로젝트 가운데 하나이며, 그 모습은 오늘날의 소셜 미디어 및 전자 상거래 콘텐츠 추천 시스템에서도 그와 유사한 형태를 볼 수 있습니다.

Netflix 및 Amazon 콘텐츠 추천

한편 1998년에 스타트업으로 설립되었던 Netflix는 우편으로 DVD를 보내주는 구독 서비스를 제공하며 Blockbuster가 주도하던 개인 비디오 대여 시장에 도전했습니다. 2000년에 Netflix는 사업 모델을 '무제한 대여' 구독 시스템으로 비즈니스 모델을 전환하여 훨씬 더 많은 콘텐츠 보유량을 확대하고 콘텐츠 보유량을 대폭 늘렸습니다. 또한 회원의 영화 평가란과 협업 필터링을 바탕으로 한 개인 맞춤형 추천 서비스 Cinematch를 추가했습니다. 2001년에는 별점을 도입하여 추천을 더욱 강화했으며, 이후 다른 알고리즘 에코시스템을 도입하여 영화 대여 프로세스를 개인 맞춤화하고 상품화했습니다.

Amazon도 이 시기에 콘텐츠 추천 실험을 실행했습니다. 이 회사는 1997년 중반에 Greg Linden의 주도 하에 잘 알려진 "웹 사이트 분할" 프로젝트를 실시했습니다. 이는 향후 클라우드 컴퓨팅으로 이어질 수평적 확장 및 활용을 개척했습니다. 이로 인해 고객 참여를 특징짓는 데 필요한 대규모 머신 데이터 컬렉션이 급격히 증가했습니다. Andrew Ng을 비롯한 전문가들은 이를 데이터와 머신 러닝의 '선순환'이라고 말했습니다. 2003년에 Amazon은 Greg Linden, Brent Smith, Jerry York가 집필한 "Amazon.com 추천 시스템: 항목별 협업 필터링"이라는 유명한 논문을 발행했습니다. 2003년의 이 시점에 저자들은 협업 필터링, 클러스터링 모델 (예: 차원 축소), 검색이라는 콘텐츠 발견 및 추천이라는 세 가지 공통 접근법을 발견했습니다. 이들의 연구는 항목별 협업 필터링을 도입했다는 점에서 차별화됐으며, 이는 "이 책을 구매한 고객은 다음 책도 구매했습니다"라는 유명한 문구로 이어졌습니다. 즉 Amazon의 추천 시스템은 사용자의 수나 콘텐츠 항목의 수와 독립적으로 확장될 수 있었다는 뜻입니다. Amazon은 유선 추천, 방대한 데이터 세트, 추천의 성공을 평가하기 위한 품질 측정 항목을 생성하기 위해 빠른 추론을 우선시했습니다.

Amazon은 사업을 확장하고 더 많은 비즈니스 라인을 구축하면서 추천 기반 전자 상거래 분야에서 선두를 달렸습니다. 이러한 노력의 직접적인 결과는 클라우드 컴퓨팅의 기원, 빅 데이터 활용 사례, 머신 러닝의 상업적 사용으로 이어졌습니다.

추천 시스템 및 소셜 네트워크

여러 소셜 네트워크가 뒤를 이어 유사한 접근법을 적용하면서 차세대 커뮤니케이션 플랫폼이 구축되었습니다. 그중에서 LinkedIn은 2006년 Jonathan Goldman이 회사에 합류하면서 규모를 확장했습니다. 네트워크 분석의 가치를 인식한 Goldman은 제품 팀의 무관심에도 굴하지 않고 "알 수도 있는 사람" (PYMK) 이라는 추천 엔진을 구축했습니다. 이는 LinkedIn 사용자들 사이에서 굉장한 인기를 얻었습니다. PYMK 기반 콘텐츠의 클릭률 (CTR) 은 초기에 다른 접근법보다 30% 높았으며, 이 비즈니스 지향 소셜 네트워크에서 추천은 빠르게 사용자 경험의 필수 구성 요소가 되었습니다.

2000년대 후반이 되면서 실제 제품에서 추천 시스템의 인기가 상대적으로 일반화되었습니다. 온라인 데이팅 시스템, 온라인 게이밍, 소매 등 다른 사업 분야에서도 이 기술은 많이 활용되었습니다. 주류 기업들이 추천 시스템을 도입하고 우수 인재가 투입되기 시작하면서 데이터 사이언스는 2009년에 일반적으로 쓰이는 용어가 되었습니다. 추천 시스템이 전자 상거래에서 증명한 가치가 주류 기업의 발전에 큰 역할을 한 것으로 보입니다.

Netflix Prize

2006년부터 2009년까지 Netflix는 Netflix Prize라는 추천 시스템 대회를 열었습니다. 여러 팀들이 식별 불가능한 고객의 영화 평점을 사용하여 추천을 최적화하는 경합을 벌인 것입니다. 대회에는 Netflix의 알고리즘을 한참 뛰어넘는 기술이 많이 출전했지만, 여러 이유로 이러한 기술은 실제 제품으로 상품화되지는 않았습니다. 그럼에도 불구하고 이는 대회에서 우수한 팀들을 만나 추천 시스템을 구축하고, 평점 데이터의 빈틈을 분석하는 방법을 학습하고, 궁극적으로 머신 러닝 작업을 실제 제품에 적용하는 것이 얼마나 중요한지 확인할 수 있었습니다.

추천 시스템의 진화

2년도 채 되지 않아 추천 시스템은 연구원의 받은 메일함을 관리하는 데 도움이 되는 가벼운 프로젝트에서 선도적인 기술 기업의 수익을 창출하는 핵심 기술로 발전했습니다. 이러한 초기 프로젝트를 통해 식별된 추천 시스템의 주요 개념 및 요구 사항은 다음과 같습니다.

- 협업 필터링
- 개인화
- 소량의 특수 데이터가 아닌 방대한 데이터 세트 사용
- 규모를 위한 분산 시스템 활용
- 사용자 또는 항목 수와 독립적으로 확장
- 동일한 사용 사례 내에서의 사용자 모델링, 콘텐츠 모델링 및 기타 머신 러닝
- 추천 품질의 정량적 평가
- 보안 및 데이터 개인 정보 보호 문제
- 빠른 추론의 필요성
- 좋은 추천 시스템 구현의 어려움
- 추천 품질을 비즈니스 매출과 연결
- 성공 지표 (예: LinkedIn의 PYMK 도입으로 인한 CTR)

관찰

역사의 시작은 10년 전으로 거슬러 올라갑니다. 추천 시스템의 역사에서 발생한 기회, 과제, 탐색에 대한 핵심 포인트는 향후 현재 트렌드를 검토할 때 활용할 것입니다.

> 기회: 사용자 참여

첫 번째 포인트는 기회에 관한 것입니다. 사용자 커뮤니티가 온라인 활동에 참여하는 경우, 해당 사용자들의 명시적 피드백과 묵시적 데이터를 일반화하여 사용자 경험을 향상화할 수 있습니다. 추천 시스템은 전자 상거래, 온라인 데이팅, 뉴스 리더, 영업 리드, 게이밍, 음악 앱 등 다양한 활동에 걸쳐 사용자 경험을 향상하는 등 여러 가지 방식으로 사용되었습니다. 거시적으로 보면 문제는 항상 발견과 관련되어있으며 이커머스 시나리오가 가장 보편적인 방식을 보여줍니다.

1. **탐색:** 고객이 구매해야 하는 위젯을 정확히 알고 있는 경우, 해당 위젯 페이지로 이동한 다음 결제하도록 돕습니다. 실제로는 이러한 '풀 (pull)' 조건은 상대적으로 드물지만 UX 디자인 최적화에는 중요합니다.
2. **검색:** 고객이 구매해야 하는 위젯을 대략적으로 알고 있는 경우, 유사한 항목을 검색하고 결과 모음 중에서 선택하도록 돕습니다. 물론 검색은 널리 알려져 있으며, 기존 사례를 통해 증명되었듯이 보다 간단하게 UX로 구현하고 도입할 수 있습니다.
3. **추천 시스템:** 고객이 다른 활동을 탐색하거나 활동에 참여한 경우, 고객의 웹 브라우징 경험을 개인화하여 추천 항목을 고객에게 '푸시 (push)' 합니다. 이 고객 시나리오가 통계적으로는 더 일반적이지만, 이를 효과적으로 구현하는 것은 이전 두 UX 시나리오에 비해 더 복잡하며 고려할 사항도 많습니다.

예시로 Netflix 비즈니스 모델이 어떻게 발전했는지를 살펴보겠습니다. 이 회사의 UX는 다음과 같은 단계를 거쳤습니다. 1999년 Netflix UX는 탐색 시스템에 약간의 검색 기능이 더해진 것이었습니다. 그 후 곧 추천과 맞춤화가 도입됐습니다. 보다 고유한 추천 시스템 사용 사례에 힘입어, 수년에 걸쳐 점점 높은 수준의 맞춤화가 추가되었습니다. 처음에는 시간이 걸렸습니다. 초기에는 Netflix에서 선택된 콘텐츠의 20%가 추천 콘텐츠였지만 나중에는 80%까지 증가했습니다.

이러한 증가세는 조직이 AI 애플리케이션을 채택하여 온라인 비즈니스 활동을 강화하기 시작할 때 나타납니다. 단순하게 시작해서 훨씬 복잡하고 정교한 접근법을 단계별로 추가하는 것입니다. 궁극적으로는 특정 사용 사례에 여러 개의 추천 시스템이 사용됩니다. 이러한 과정에서 추천에 대한 의존도는 사용자 경험의 '일부'에서 '대부분'으로 변하게 됩니다.

> 과제: 추천 품질

두 번째 포인트는 과제에 관한 것입니다: 적절한 추천 시스템을 효과적으로 작동시키는 것은 매우 어렵습니다. 모델 훈련에 사용할 수 있는 좋은 데이터는 얻기 어렵습니다. 사람들이 선호하는 것은 극적으로 바뀌곤 합니다. 예를 들어 보겠습니다. 가치가 높은 광고 카테고리 중 하나는 고급 자동차입니다. 새로운 고급 자동차 판매에 대한 추천은 매수/매도 광고 네트워크에서 잘 작동하는 경향이 있습니다. 다시 말해 새 자동차를 구매할 의향이 있는 사람은 관련 광고를 클릭하는 경향이 있고, 이는 가치로 이어집니다. 그러나 개인이 새 차를 구매하고 나면 관련 광고를 클릭할 일은 없어지게 되죠. 한동안 지속적으로 좋은 추천이었던 것이 갑자기 나쁜 추천이 됩니다. 누가 무엇을 언제 구매했는지에 대한 정보가 부족하기 때문에 광고 추천 시스템에는 품질에 대한 내재적인 위험이 있으며, 이 문제는 추천 시스템의 응용 영역 대부분으로 확장됩니다.

추천 품질에 대한 불만은 1979년 Elaine Rich의 Grundy 시스템까지 거슬러 올라가며, 그 이후로 계속되어 왔습니다. 특정 시점에 특정 사용자에게 일부 항목을 추천하는 것은 가능하지만, 특정 사용자가 원하는 항목을 정확하게 추천하는 동시에 원하지 않는 항목을 피하는 것은 매우 어렵습니다. 그럼에도 이는 사용자가 자동화된 시스템을 평가하는 일반적인 판단 기준입니다.

게다가 일종의 양날의 검과 같이, 추천 시스템이 지나치게 좋은 경우 사용자 경험을 불쾌하게 만듭니다. 미학과 디자인에서 불쾌한 골짜기는 자동화가 인간을 잠식하는 것으로 여겨질 경우 관찰자들 사이에 불쾌함을 유발한다는 개념입니다. 이는 좋은 추천 시스템이 사용자에 대해 '지나치게 많이' 알고 있다고 느끼게하는 딜레마를 제기합니다. 추천 시스템에서 개인 정보 보호에 대한 우려 사항은 1992년 Doug Terry의 Xerox PARC로 거슬러 올라갑니다.

물론 머신 러닝은 추천 시스템의 역사에서 중요한 역할을 수행했으며, 머신 러닝 분야 또한 같은 시기에 빠르게 발전했습니다. 머신 러닝은 과거 데이터 또는 합성 데이터를 활용하여 패턴을 생성한 다음 다른 맥락에 적용하는 것입니다. 이는 현실에서 추천 시스템에 필요한 사항과 완벽하게 들어맞습니다. 중요한 것은 머신 러닝 모델이 훈련 데이터를 일반화하면서 맥락을 '놓친다'는 사실입니다. 특정 사용자에 대한 특정 추천 집합이 만들어진 이유를 설명하기 어려워지는 것이죠. 이는 데이터 팀이 머신 러닝 모델을 활용한 추천 품질이나 문제 해결 워크플로우를 작업해야 할 때 심각한 문제를 야기할 수 있습니다.

> 내러티브: 추천 시스템 사용 모범 사례

세 번째 포인트는 내러티브에 관한 것입니다. 이 초기 성장 기간 동안 추천 기술에 대한 일반적인 인식은 소수 기업들의 내러티브를 통해 확산됐습니다. 대규모 사용 사례에 대한 내러티브는 기술적인 관점에서는 흥미롭고 종종 업계 분석가들에 의해 '모범 사례'라고 여겨지기도 했지만, 항상 일반적으로 아주 훌륭한 추천 시스템 사용 사례라고 볼 수만은 없었습니다.

초기 연구는 추천을 정보 검색의 관점에서 바라봤고, 모델은 정확도와 검색 결과로만 평가되었습니다. 그 결과 Elaine Rich의 사용자 모델링이 보여줬던 깊이는 잊혔습니다. 또한 이러한 학술적인 관점은 복잡한 대규모 생산 시스템에서 문제를 해결하고 미세조정하는 데 특별히 효과적이지 않았습니다.

Netflix 및 유사 사례의 상업적 성공은 이후 비음수행렬분해 (Non-negative Matrix) 의 활용을 부각시켰습니다. 이 내러티브는 Apache Hadoop과 같은 빅 데이터 도구의 인기와 맞아 떨어졌습니다. 최근에는 딥 러닝이 인기를 끌면서 추천 시스템의 활용은 신경망과 임베딩 (Embedding) 으로 바뀌었습니다.

이러한 내러티브와 트렌드는 선도적인 기술 기업의 혁신을 더 많은 청중에게 전파하는 데 도움이 됩니다. 또한 점점 더 복잡해지는 문제를 보다 정교한 접근법으로 해결하려는 수요가 증가함에 따라, 내러티브는 데이터 사이언스 전반에 대한 사고의 진화를 추적하는 데도 도움이 됩니다. 그러나 이러한 유형의 내러티브는 좋은 추천 시스템 관행을 수립하는 근본 동인과 트레이드오프 간의 경계를 흐려지게 만들기도 합니다. 다음 섹션에서 이러한 동인과 트레이드오프에 대해서 자세히 알아보죠.

전반적으로 추천 시스템의 결과가 효과적이려면 데이터 팀이 여러 경쟁 우선 순위와 우려상황 속에서 균형을 맞춰야 합니다. 이는 간단한 작업이 아닙니다. 추천 시스템 활용에서는 이러한 절종류의 트레이드오프 사이에 알력이 존재하며, 그중 일부는 향후 계속해서 문제가 될 것입니다. 일부 문제는 더 나은 도구를 통해 해결할 수 있습니다. 다음으로 업계의 추천 시스템 팀이 실제 환경에서 어떤 방식을 취하고 있는지 알아보겠습니다.

인터뷰: 주요 내용 발췌

저희는 추천 시스템의 전문가로 엄선된 팀 및 개인과 협업하여 추천 시스템 워크플로를 담당하는 선임 기술자들을 인터뷰했습니다. 간결함을 위해 다음 섹션에서는 전체 인터뷰에서 핵심 사항만을 발췌했습니다.

각 인터뷰에서는 팀의 이력, 팀의 초점 및 책임, 팀 교육 방식에 대해 질문했습니다. 데이터 속도와 규모, 데이터 준비 및 특징 엔지니어링에 사용되는 접근 방식, 적절한 기술 및 프레임워크를 선택하는 방법, 모델을 평가하고 튜닝하는 방법, 추천 시스템 최적화를 위한 작업 등 추천 시스템 활용에 대해 구체적으로 물었습니다. 또한 추천 시스템을 시작하려는 사람들을 위한 일반적인 조언도 요청했습니다.

이 인터뷰 내용들의 공통점을 알아보고 업계 내의 여러 활용 간 차이점을 찾아보겠습니다. 가능한 경우 추천 시스템 역사의 주요 내용을 분석에 활용할 것입니다.

Monica Rogati, 독립 고문

AI 및 데이터 사이언스 고문, LinkedIn PYMK의 첫 머신 러닝 모델 개발자

Monica Rogati는 CMU에서 컴퓨터 공학 박사 학위를 받았으며, 응용 머신 러닝, 자연어 처리 (NLP), 웨어러블 기기 등의 분야에서 경력을 쌓았습니다. 그녀는 LinkedIn 데이터 사이언스 팀의 초기 멤버로서 추천 시스템, 기타 데이터 제품을 개발했으며 이후 Jawbone의 데이터 담당 부사장이 되어 데이터 사이언티스트 및 엔지니어로 구성된 전문가 팀을 조직하고 이끌었습니다. 현재 독립 고문인 Monica는 "AI 욕구 단계" 다이어그램의 창안자로 유명합니다. 이 다이어그램은 오해 받고 있는 데이터 사이언티스트들을 보호하는 역할을 합니다.

추천자의 규모와 데이터 속도 측면에서 Monica는 제로 데이터부터 수백만 명의 사용자를 수많은 항목과 연결하는 시스템까지 다양한 시스템을 구축했습니다. Monica는 "이 분야의 최상위 기술은 롱테일을 다루고 마법처럼 (또는 관점에 따라서는 불쾌하게) 굉장히 개인화된 추천을 제공하는 만큼 매우 주목을 받습니다."

> 추천 시스템의 주요 과제

현실에서 추천 시스템의 주요 과제는 무엇입니까? 가장 큰 과제 중 하나는 지정된 사용 사례에 적합한 메스릭스 및 목표 함수를 결정하는 것입니다. **"데이터 사이언티스트는 결국 '메트릭스 엔지니어'가 될 것입니다. 메트릭스은 자동화가 가장 어렵고, 거의 불가능에 가까운 데이터 작업 중 하나입니다."**

Monica는 실무자가 추천 시스템의 목표 함수, 즉 극대화하려는 메트릭스에 대해 신중하게 생각해야 한다고 경고합니다. 이는 사용 사례가 달성하고자 하는 것의 프록시를 나타냅니다. "당신의 시스템에 영향을 미친 다음에도 좋은 프록시가 될 수 있을까? 인센티브 일치로 인해 자발적이든 우발적이든 조절할 방법은 있는가? 문제는 얼마나 어려운가? 그리고 성능에서 기대할 수 있는 상한선은 무엇인가? 예를 들어 학습 데이터의 채점자 간 일치도는 얼마나 되는가?와 같은 문제를 고려해야 합니다."

Monica는 또한 추천 시스템을 구축하기 위해 적절한 기술을 선택하는 법에 대해 조언했습니다. 문제 해결 가능성 외에도 통합의 용이성, 성숙도, 선택권 보존 등 고려해야 할 몇 가지 기준이 있습니다. 가장 중요한 기준은 회사의 현재 워크플로우와의 통합 용이성과 결과 대비 투입 리소스의 비율을 최소화하는 것입니다.

"지난 몇 년 동안 데이터 도구와 프레임워크가 폭발적으로 증가했습니다. 대다수 도구가 특정 환경에서 입증되고 신뢰할 수 있을 때까지는 워크플로에 큰 변화 없이도 쉽게 통합될 수 있게 됐다는 점은 좋은 일입니다. 이런 도구를 만드는 분들에게 제가 항상 묻는 질문은 이렇습니다. '여러분의 트로이 목마는 무엇인가요? 채택을 받고 신뢰를 얻기 위해 실무자의 노력을 거의 또는 전혀 들이지 않고도 해결 가능한 문제는 무엇인가요?'"

중요한 점은 개발 수명 주기 초기에 기술 선택에서 위험을 제거하고 비상시를 위한 백업 계획을 수립하는 것입니다. "최고의 도구는 마케팅의 잠재적 탈출구를 포함하며 쉬운 통합 지점뿐 아니라 데이터 내보내기 도구 또한 제공합니다." 이는 오픈 소스에 적합한 경향이 있습니다. '통합 용이성'을 우선시되는 경우, 하나의 컴포넌트를 교체하여 혼란을 억제하여 '선택지 보존'을 허용하는 동시에 성과 대비 노력의 비율을 유지할 수 있습니다.

> 추천 프로젝트 시작하기

"IBM Research에서 인턴으로 일하면서 Salim Roukos에게 배운 것이 있습니다. 머신 러닝 시스템으로 작업을 시작하기 전에 1시간 동안 데이터 레이블링과 알고리즘의 작업을 수행하라는 것입니다. 알고리즘 그 자체가 되세요. 이 일련의 작업을 통해 수많은 데이터 품질 문제를 발견하거나, 질문이 잘못 정의됐을 때의 상황들을 식별하거나, 언제 추천을 하지 말아야 하는지 등을 알 수 있습니다.

일반적으로 모델을 평가하기 위해 Monica Rogati는 위험 제거 프로젝트에 도움이 되는 3단계 프로세스를 다음과 같이 설명합니다.

1. 간단한 기준선 만들기
2. 넉넉한 오프라인 필터 구축
3. 느리게 증가하는 A/B 테스트

첫 단계는 데이터를 분석한 다음 합리적이면서 간단한 규칙 기반 기준선을 만드는 것입니다. 이미 모델이 있는 경우에도 이를 시도해 보세요. 간단한 기준선은 디버깅이 용이하고, 견고하고, 빠릅니다. 현재 모델만큼 좋은 경우 이를 우선해야 합니다.

'관대한 (lenient) 오프라인 필터' 사용은 기존의 오프라인 모델 평가 개념을 바꿔 놓습니다. 모델이 큰 혼란이나 예기치 못한 피해를 일으키는지, 그 결과가 현실에서 대략적으로 유망해 보이는지 확인하세요. 반드시 기존 시스템을 능가하는 것이 목표는 아닙니다. 그러나 간단한 기준선 위에서 시작했다면 그래야 합니다.

마지막으로 실제 가동 환경에서 A/B 테스트를 사용하여 훈련한 머신 러닝 모델을 매우 느리게 평가합니다. 예를 들어 1%, 2%, 5%, 10%, 20% 등 각 단계에서 메트릭스를 확인합니다. 운영 문제, 누락된 기능, 예기치 않은 피해, 기타 일반적으로 오프라인 또는 소규모에서는 나타나지 않을 수도 있는 모든 문제를 확인합니다.

머신 러닝 모델이 본격적으로 실행된 후, 특히 추천 시스템의 경우 이상치 효과가 사라질 수 있도록 'Burn in' 기간이 필요한 경우가 많습니다.

Xiangting Kong, Tencent

Tencent의 전문 엔지니어, 광고 추천 시스템 설계 및 개발 담당, Tencent 광고 및 딥 러닝 플랫폼 리드

Xiangting Kong은 광고 교육 플랫폼의 최적화를 담당하는 광고 추천 시스템 전담 Tencent 팀을 이끌고 있습니다. 이 플랫폼의 구성 요소로는 오프라인 특징 엔지니어링, 훈련 플랫폼, 온라인 추론 시스템, 온라인 특징 엔지니어링 및 플레이 플랫폼 등이 있습니다. 이들은 광고부터 핀테크, 네트워크 데이터 마이닝 등 다양한 비즈니스 시나리오에서 모델 훈련, 최적화, 추론 작업을 진행합니다. 또한 관련된 기술 및 업무의 범위를 고려하여 팀 교육을 위해 1~2주마다 기술 공유 세션을 구성합니다.

Xiangting은 다음과 같이 조언합니다. “광고 추천은 점진적인 필터링 과정입니다. 정렬 단계에는 리콜, 사전 순위 지정 및 순위 지정이 포함됩니다. 각 단계마다 요구 사항이 다릅니다. 모델의 신속한 조사와 반복으로 인해 훈련 성과에 대한 요구 사항은 더욱 높아졌습니다."

데이터 속도 및 모델 복잡성에 대해 고려해야 할 중요한 절충 요소가 있습니다. 교육/테스트 데이터의 양을 늘리고 특징 세트를 복잡하게 만들면 모델 품질을 개선할 수 있습니다. 한편, 모델 훈련 시간이 증가하고 모델 업데이트 속도가 제한되기도 합니다. **"샘플 특징을 추가하여 더 많은 샘플 데이터를 훈련시키면 광고 추천의 정확도를 높일 수 있습니다. 그러나 이는 훈련 시간을 늘리고 모델의 업데이트 빈도에 영향을 미칩니다."**

Tencent 팀은 추천 훈련 프레임워크에 HugeCTR을 사용합니다. 이것은 광고 추천 플랫폼에 통합되어 모델 교육 및 업데이트 빈도를 가속화합니다. 또한 훈련 중에 더 많은 데이터 샘플을 사용하여 추천 결과를 개선할 수 있습니다. 일반적인 기술 선택의 관점에서 Xiangting은 오픈 소스 프로젝트 작업이 우선시되는 이유를 다음과 같이 설명합니다. "선택한 기술이나 프레임워크가 커뮤니티 생태계와 호환되어야 더 나은 후속 업그레이드를 수행할 수 있습니다."

> 최근 연구 활용

현재 활용되고 있는 연구를 살펴보면, Tencent는 최근 압축 희소 행렬 (CSR) 파이프라인을 자사의 광고 추천 시스템 훈련에 도입했습니다. "CSR 방식의 훈련 데이터가 생성되기 때문에 훈련용 GPU에서 직접 데이터를 읽을 수 있습니다. 데이터 처리 파이프라인 최적화를 통해 CPU 로드가 크게 줄었고 GPU 활용은 대폭 향상됐습니다.”

Xiangting은 추천자 시스템 여정을 막 시작하는 팀들을 위해 빠르게 발전하는 환경 속에서 꾸준한 경로를 유지할 것을 조언합니다. "성숙한 기술 프레임워크를 선택하고 커뮤니티 생태계와 호환을 유지하여 후속 시스템 업그레이드를 용이하게 하세요."

Felipe Contratres, Magalu

Magalue의 개인화 팀 리더 (Magazine Luiza)

Felipe Contratres는 추천 및 검색 플랫폼을 운영하는 Magalu의 개인화 팀을 이끌며 자사 플랫폼과 새로운 추천 모델을 위한 새로운 기능을 개발합니다. 추천 시스템 개발을 막 시작하는 사람들을 위해 Felipe는 다음과 같이 조언합니다. "간단한 접근법으로 시작하세요. 다수의 문제는 전체 시스템 복잡성을 낮추는 간단한 기법으로 쉽게 해결됩니다. 고급 기술이 필요한 경우 워크플로우에 통합할 성숙한 패키지 또는 프레임워크를 선택하세요." 또한 Felipe는 가능한 한 빨리 프로덕션 환경에서 새로운 솔루션을 테스트할 것을 권장합니다. "오프라인에서 과도한 사전 최적화를 진행하는 대신, 실제 고객과 함께 가능한 빨리 온라인으로 모델을 테스트하세요."

팀은 테이블 형식의 이벤트를 텍스트 및 이미지 벡터와 결합하기 위한 다중 모달 기능을 찾고 있습니다. "최근 논문에서 Gabriel de Souza P. Moreira가 제안한 기술을 테스트하는 것을 고려하고 있습니다." 또한 세션 간에 사용자 환경을 사용자 지정하는 경험적 접근 세션 기반 추천 시스템도 있습니다. 이 세션 기반 추천 시스템은 이전 연구들을 Magalu의 비즈니스 지식과 결합하여 개발되었습니다.

Jun Huang, Meituan

Meituan 머신 러닝 플랫폼의 훈련 프레임워크 팀을 담당하는 Meituan의 수석 기술 전문가.

Meituan에 따르면 "Meituan의 사명은 사람들이 더 잘 먹고, 더 잘 살도록 돕는 것"입니다. 중국의 선도적 전자 상거래 서비스 플랫폼으로서 Meituan의 비즈니스는 '음식 + 플랫폼' 전략을 중심으로 구축되어 있으며 '식사'를 핵심에 두고 있습니다. Meituan은 Meituan, Dianping, Meituan Waimai 등 다수의 유명한 모바일 앱을 운영합니다. Meituan의 사업 범위는 매우 광범위하여, 중국 전역 2,800개 이상의 도시에서 케이터링, 온디맨드 배송, 자동차 운송, 자전거 공유, 호텔 및 여행 예약, 영화 발권 및 기타 엔터테인먼트 및 라이프스타일 서비스 등 200가지 이상의 서비스 카테고리를 운영합니다.

Meituan의 훈련 프레임워크 팀은 대규모 CPU/GPU 클러스터에 배포된 고성능 분산 딥 러닝 트레이닝 프레임워크를 개발합니다. 추천 시스템에서는 최대 1000억 개의 희소 파라미터와 1000억 개의 샘플로 분산 학습을 지원합니다. Jun은 "최근에는 NVIDIA A100을 기반으로 차세대 추천 훈련 시스템을 설계해 교육 효율성과 모델 복잡성을 크게 향상시켰습니다"라고 설명했습니다.

적절한 기술을 선택하는 방법은 무엇이느냐에 대한 질문에 Jun은 오픈 소스를 우선시한다고 답했습니다. "진보적이고 개방적이며, 생태 호환적이어서 이를 바탕으로 내부의 요구를 더 잘 충족할 수 있는 기술이어야 합니다. 저희 팀은 현재 오픈 소스 기술을 주로 사용하여 시스템을 구축하고 있습니다. 그와 동시에 저희의 작업 결과로 오픈 소스 커뮤니티에 기여할 수 있어서 기쁩니다."

하드웨어는 전략에서 중요한 역할을 합니다. Jun은 "처음에는 CPU 아키텍처를 기반으로 교육 프레임워크를 최적화했지만, 모델이 점점 더 복잡해짐에 따라 교육 프레임워크를 심층 최적화하기가 어려웠습니다. 지금은 A100 GPU를 기반으로 훈련 시스템에 NVIDIA HugeCTR을 통합하려고 작업 중입니다. 8개의 A100 GPU가 있는 단일 서버는 CPU 기반 훈련 시스템으로 일하는 수백 명의 근로자를 대체할 수 있습니다."라고 밝혔습니다.

이제 막 추천 시스템을 시작하는 사람들에게 Jun은 다음과 같이 조언합니다. "현재 회사의 인프라와 비즈니스 상태를 충분히 이해하고 이를 바탕으로 시스템 및 프로세스를 설계하세요. 기술 스택과 프레임워크를 선택할 때 각 시스템의 성숙도, 커뮤니티 생태, 확장성 및 통합 친화성을 고려해야 합니다."

Chris Wiggins, The New York Times

New York Times 수석 데이터 사이언티스트

Chris Wiggins는 The New York Times의 데이터 과학 팀을 이끌며, 보도국 및 비즈니스 문제를 위한 머신 러닝 솔루션을 개발하고 배포하고 있습니다.

추천 시스템이 최신 뉴스 발행에서 수행하는 역할에 대해 Chris는 다음과 같이 설명합니다. "구독자들은 The Times 편집부의 판단에 관심이 많습니다. 따라서 **추천은 신뢰와 충성도에 맞춰 최적화하여 편집부의 판단을 대체하지 않고 강화하는 방향으로 일하도록 만드는 데 그 의미가 있습니다."** 추천 시스템은 "편집자의 선택", "가장 인기 있는 기사", "더 스마트한 생활", 요리 앱, 모바일 앱의 "당신을 위한 기사" 개인 맞춤형 탭, "당신을 위한 주간 에디션" 뉴스레터 등 다양한 뉴스 서비스를 강화하는 데 활용됩니다.

알맞은 기술 선택에 대해 Chris는 Kreps 법칙을 활용한다고 설명했습니다. Apache Kafka 공동 개발자인 Jay Kreps의 이름을 딴 법칙입니다. "제품 개발 연구를 위한 팁: 3~5개의 펍 (pub) 을 읽고 그것들이 극복하겠다고 주장하는 단순한 것들에 주목한 다음, 실행하세요."

그러나 그와 반대로, “전술은 제품과 뉴스룸 파트너의 목적에 달려 있습니다. 뉴스룸 파트너는 ‘x’ (맥락), ‘a’ (어떤 활동, 기사, 자산을 추천할 것인지), ‘y’ (어떤 결과를 최적화하고자 하는지) 를 깊이 생각합니다.

The New York Times의 추천은 Samizdat라는 플랫폼을 통해 제공됩니다. 이 서비스는 확장과 개인 정보/규제 준수를 모두 처리합니다. 본질적으로 플랫폼은 독자에 대한 정보를 수집하고, 관련된 개인 정보 규제에 따라 이를 분석한 다음 독자의 상호작용을 어떻게 처리할지에 대한 지침을 출력합니다. 이를 통해 인프라 팀은 여러 지역에서 데이터 규정 해석을 조정하고 전체 제품군에서 변경 내용을 간소화할 수 있습니다.

이러한 활용으로 많은 추천 모델을 관리합니다.

추천 시스템 여정을 막 시작하는 새로운 팀을 위해 Chris는 다음과 같이 조언합니다. “저는 Kreps 법칙이나 Monica Rogati의 ‘AI 욕구 법칙’ 인포그래픽 등을 활용하는 편입니다. NYT의 성공 비결은 데이터 사이언티스트와 소프트웨어 엔지니어가 나란히 앉아 함께 개발하고, Python (데이터 사이언티스트) 과 Go (소프트웨어 엔지니어링 파트너) 를 혼합하여 코딩함으로써 ‘통계와 SLA’ 양쪽에서 성과를 내는 메서드를 구축한 것입니다.”

> 추천 시스템의 주요 과제

대상 그룹과의 신뢰 구축, 콘텐츠 측면에서 보다 깊이 있고 세분화된 모델 등 NYT의 우선 순위는 잘 작동하는 머신 러닝 모델이 어떤 것인지 보여줍니다. “한 가지 성공은 예측 방법에서 규범적 방법으로 전환한 것이었습니다. 예를 들면 다양한 상황별 밴디트 (bandits) 를 콘텐츠의 서로 다른 표면 및 풀에 대해 배포할 수 있습니다.”

추천 시스템을 통한 개인화는 편집부의 판단을 보완하며 뉴스 보도에서 어떤 기사를 헤드라인으로 내보낼지 결정합니다. "The Times의 알고리즘 큐레이션은 웹 사이트 및 앱의 지정된 부분에서 사용됩니다. 홈페이지의 '더 스마트한 생활' 섹션이나 개인 뉴스레터인 '당신의 주간 에디션'과 같이 수동 큐레이션이 비효율적이거나 어려운 곳에서 콘텐츠를 선택하는 데 사용합니다."

Chris는 다음과 같이 설명합니다. "기사 추천에 상황별 밴디트를 사용하면 변화하는 기본 설정에 신속하게 적응하고 새로운 옵션을 효율적으로 탐색하는 데 유용합니다.” 물론 현실에서 제품의 정교한 추천 시스템에는 굉장히 많은 단계가 있을 것이므로, 상황별 밴디트는 추천 시스템 워크플로 구성을 위한 툴키트의 구성 요소로 볼 수 있습니다. 특히 개인화로부터 업스트림으로 발생하는 콘텐츠 모델의 경우 그렇습니다. "간단한 상황별 밴디트를 구축하면 필요한 성능을 확보할 수 있습니다. 콘텐츠가 쏟아져 나오는 뉴스 비즈니스에서는 예측 가능성과 이해를 위한 해석 가능성이 균형을 이뤄야 합니다.”

Chris는 향후 전망에 대해 다음과 같이 말했습니다. “저희는 항상 새로운 고지를 바라보며 뉴스 지형의 변화에 대응하려고 노력하고 있습니다.” Chris는 자신의 팀이 추천 시스템 여정을 기록해 나가는 https://open.nytimes.com에 주목할 것을 권장했습니다.

Kannan Achan, Walmart Global Tech

Walmart Global Tech에서 개인화 및 추천 팀 지도

Kannan Achan은 Walmart Global Tech에서 고객 여정을 개인화하는 팀을 이끌고 있습니다. “저희는 주어진 페이지에 대해 전체 페이지 개인화를 수행합니다. 이는 크리에이티브한 배너, 상품 진열대 혹은 CAD가 될 수도 있습니다. 모든 것은 개인화 팀에 한 번에 주어지며, 전체 페이지를 한 번에 개인화합니다.”

이 팀은 고객 이해, 콘텐츠 이해, 온라인 추론 등 세 개의 그룹으로 조직되어 있습니다. 후자는 런타임에서 수행된 제품 간 고객 특징 및 콘텐츠 특징과 관련되어 있습니다. 이렇게 관심 분야를 분할하면 조직에 유연성이 생긴다고 Kannan은 설명합니다. “이는 확장하고, 협업하고, 흥미로운 문제 다수를 해결하는 데 큰 도움이 됩니다.

추천 시스템은 Walmart 웹 사이트 및 Walmart 매장에 걸쳐 옴니채널 데이터를 사용합니다. 추천 시스템 사용 사례 과정을 살펴보면, 초기에 옴니채널 데이터는 분석에 사용됐습니다. 이후 회사는 고객에 대한 보다 풍부한 이해를 필요로 하게 됐습니다. 이는 개인 정보 보호에 대한 충분한 고려와 균형을 이뤄야 했습니다. “저희는 이 모든 곳에서 많은 고객들이 쇼핑을 하는 모습을 보고 있습니다.” 옴니채널 데이터 시나리오를 복잡하게 하는 요소 중 하나는 고객 데이터 내에 의도라는 측면이 있다는 것입니다. 이는 온라인과 오프라인 매장 양쪽의 많은 소매 맥락에서 마주치게 되는 요소입니다. 예를 들어 고객이 주간 식료품 쇼핑을 하고 있을 수도 있지만, 또 어떤 때는 그 동일한 고객이 TV와 같은 일회성 구매를 고려할 수도 잇습니다. 추천 시스템의 활용은 의도에 따라 굉장히 달라질 수 있습니다. “추천 시스템 부문을 생각해 보면, 기존의 관점에서 추천 시스템은 사용자와 평점으로 구성된 행렬입니다. 즉 협업적 필터죠.” 발견을 일반적으로 살펴보면, ‘이 상품을 둘러본 고객은 다음과 같은 상품도 둘러보았습니다’ 같은 전통적인 추천을 제공해야 합니다. 그러나 상품이 장바구니에 들어가고 나면 이를 보완하는 상품 또한 추천해야 하겠죠.”

> 추천 프로젝트 시작하기

Monica Rogati의 조언과 마찬가지로 Kannan은 실무자들이 주어진 사용 사례에 대한 목적 함수를 생각하고 제대로 이해하는 데 시간을 들여야 한다고 조언합니다. 계절성과 같은 복잡성을 고려하면 데이터의 맥락을 이해하는 것이 더욱 중요합니다. "예를 들어 텍사스에서는 토마토를 구입하는 사람들이 할라피뇨도 같이 구입하지만, 다른 곳에서는 그렇지 않습니다. 유리병을 사서 피클을 담글 수도 있는 거죠.” 마찬가지로 “추천 시스템은 하나가 있다고 해서 모든 곳에 적용할 수 없습니다. 모든 추천 시스템은 저마다 목적 함수가 다릅니다. 이는 어느 정도 비즈니스 전략에 의해 주도됩니다.”

개념적인 관점에서 Kannan은 전체론적인 접근법을 권장합니다. “페이지 하나만 개인화하는 것으로는 부족합니다. 이를 어떤 면에서는 텐서 (tensor) 로 생각하고 전체 세션도 개인화해야 합니다.” 이 경우 고객을 이해하는 것이 최우선 순위가 됩니다. 추천 팀은 고객이 사이트에 있는 이유를 파악한 다음 고객이 자신에게 어떤 선택지가 주어졌는지 알 수 있게 해야 합니다. “발견 단계 (Discovery funnel) 에서 추천이 활용되어 전환을 돕는 것입니다. 보다 중요한 것은 이 여정에서 저희가 고객을 다시 참여하게 만드는 모델을 작업한다는 사실입니다.”

뿐만 아니라 고객이 각 참여 단계에서 어떤 행동을 하는지도 주의 깊게 고려해야 합니다. “제로 쿼리 문제는 전자 상거래에서 아주 고질적인 문제이며, 이를 해결하는 방법은 추천 시스템뿐입니다.” 제로 쿼리란 홈 페이지에서 고객의 컨텍스트밖에 없고, 고객에 대한 이해가 기껏해야 참조 URL뿐인 상황에서 전체 개인화 페이지를 만들어야 하는 경우를 말합니다.” 이에 반해 Netflix Prize에서는 추천 시스템을 포인트 솔루션으로 제시했습니다.

적절한 기술을 선택하는 방법에 대해 Kannan은 다음과 같이 말합니다. **“문제를 추정하지 마세요. 솔루션을 추정해야 합니다.”** Kannan은 또한 모델의 복잡성과 소요되는 응답 시간 간의 절충에 대해서도 설명했습니다. “40~50밀리초라는 제약 때문에 온라인에서 실행 가능한 것에는 상당한 제한이 있습니다. 이는 저희가 선택하는 기술에 큰 영향을 미칩니다. 때로는 간단한 로지스틱 회귀 분석이나 그레이디언트 부스팅 의사 결정 트리를 사용하는 것이 매우 편합니다. 실제로 효과가 있었고, 예측 가능하게 설명되었으며, 지연 시간을 매우 의미 있게 제어할 수 있었기 때문입니다."

머신 러닝 인프라의 관점에서 팀은 오픈 소스에 의존합니다. "저희는 오픈 소스에 완전히 의존합니다. 예를 들어 PySpark, TensorFlow, Airflow, Kubernetes 등을 많이 활용합니다. NVTabular, HugeCTR 등 NVIDIA의 오픈 소스는 저희가 미래의 워크플로우를 고안하는 데 상당한 영향을 미쳤습니다. 뿐만 아니라 식료품 구매를 온라인에서도 모두 가능하도록 대체해 주는 그래프 신경망 모델도 제공합니다.

> 추천 시스템의 주요 과제

하드웨어는 Walmart Global Tech의 추천 시스템에 대단히 중요합니다. "지연 시간은 매우 중요합니다. 복잡한 모델을 구축하고 온라인에서 그 점수를 매길 수는 없습니다. 보통 30~40밀리초 내에 추론할 수 있는 단순 회귀를 사용했습니다. 그러나 아주 유망한 가능성을 보여준 딥 러닝 모델이 생겼습니다. GPU를 사용하면 프로덕션 환경에서 딥 러닝 모델을 원활하게 테스트하고 확장할 수 있으며, 이는 매우 큰 차별점을 형성합니다."

팀이 면밀히 관찰하는 또 한 가지 영역은 개인 정보 보호와 공정성입니다. “저희는 이를 항상 인지하고 있습니다. 그것이 우리가 중요하게 생각하는 영역입니다.” 이 지점에서 Walmart는 The New York Times와 마찬가지 이유로 추천 시스템을 위한 상황별 밴디트에 흥미를 가졌습니다. Kannan은 다음과 같이 설명합니다. “잘 알려진 문제이기 때문에 탐색/활용에서 상당한 매력을 느끼고 있습니다. 상황별 밴디트와 보상 모델을 사용하면 큰 이익을 얻을 수 있습니다. 활용 프레임워크는 매출뿐 아니라 프리젠테이션 편향에도 도움이 됩니다. 이것이 비즈니스에 매우 유용할 수 있다는 점은 이미 수치로 증명된 바 있으며, 편향되지 않은 데이터 세트를 시도하기 때문에 저희 모델 또한 개선할 수 있습니다.”

이러한 접근법을 좀 더 자세히 살펴보면, 이는 상황별 밴디트를 워크플로우 내에 전략적으로 위치한 효율적 학습 단위, 기초 구성 요소로 활용하는 것입니다. "기존의 추천 시스템은 최상위 10개 후보만 표시했죠. 사용자는 최상위 10개 검색 결과와만 상호작용합니다. 최하위나 20위, 30위는 상위 제품이 품절되지 않는 한 결코 노출될 일이 없습니다. 그러나 모델이 탐색/활용을 수행할 경우, 특정 제품의 실적이 좋은 경우 온라인 학습을 계속하지만 분산도가 매우 높은 경우 이를 빠르게 표시하여 잘 작동하는지 확인합니다. 마찬가지로 특정 제품의 실적이 좋지 않지만 확실하지 않은 경우, 해당 제품을 더 많이 표시하여 제거할 수 있는지 확인합니다. 저희는 Thompson 샘플링의 변형을 활용하여 이를 아주 원칙에 입각한 방식으로 수행합니다.”

다른 연구 영역에서 팀은 딥 러닝된 애플리케이션으로 반복 작업을 수행합니다. “지식 임베딩 (Embedding) 과 딥 러닝된 모델을 활용한 온라인 추론은 저희가 연구 중인 주요 영역입니다.” NVIDIA Triton Inference 프레임워크는 굉장히 우아해서 CPU 기반 모델과 GPU 기반 모델 사이를 매끄럽게 전환하도록 돕습니다. 표현이 풍부하고 GPU가 작동하는 경우 ‘넓고 깊은 네트워크’를 사용하지만, 깊은 표현이 없는 경우 ‘넓은’ 네트워크를 활용합니다. 이것을 구현했다면 상당한 시간이 걸렸겠지만 NVIDIA Triton Inference 프레임워크 덕분에 구현을 가속화하기가 수월했습니다.”

Kannan은 향후 자사의 추천 시스템을 강화하기 위한 두 가지 기술 영역을 다음과 같이 언급했습니다. “유망한 연구가 많다고 여겨지는 영역 하나는 자기 지도 학습입니다. 고객이 사이트를 보면 저희는 누군가가 사이트를 봤다는 걸 압니다. 장바구니에 물건을 담는 데 10분, 결제까지 30분이 걸렸다면 그런 진행 상황도 알죠. 매장 데이터에는 ‘고객이 이 상품을 구매했다’고 기록됩니다. 하지만 그 상품을 구매한 이유, 상품 구매 순서는 알 수 없었습니다.” 이전까지 팀은 옴니채널 데이터를 서로 연관지어서 두 데이터 소스를 연계해야 했습니다. “데이터 소스 하나는 장바구니에서, 다른 하나는 명확한 의도와 그 의도가 어떻게 행동으로 전환되는지에서 나옵니다.” 자기 지도 학습은 이러한 데이터 소스에 걸쳐 고객에 대한 이해를 조화시킬 수단을 제공합니다. 다시 말해, 한 소스의 샘플링으로 다른 소스의 관련된 측면을 예측할 수 있다는 것입니다. 마치 장님 4명이 코끼리에 대한 자신의 관찰 내용을 비교하는 최적화된 버전이라고 할 수 있습니다. 생성적 적대 신경망 (GAN) 은 비지도 또는 반지도 파이프라인으로 데이터 소스 및 학습 전반을 강화하는 또 다른 방법을 제시합니다. “저희가 지난해에 투자한 영역 중 하나는 GAN 활용입니다. 특히 개인 정보 및 공격 마이그레이션 분야에서 적대 신경망으로 좋은 성과를 얻었습니다.”

Even Oldridge, NVIDIA

NVIDIA에서 NVIDIA Merlin™ 엔지니어링 리드

Even Oldridge는 NVIDIA에서 NVIDIA Merlin 엔지니어링 팀을 이끌고 있습니다. Merlin은 포괄적 GPU 가속 추천 시스템 구축을 지원하는 오픈 소스 프레임워크입니다. 데이터 전처리 및 특징 엔지니어링부터 딥 러닝 모델 훈련 및 프로덕션 환경의 추론 실행 등을 전반적으로 지원합니다.

Even은 다음과 같이 설명합니다. “Merlin은 추천 시스템에 대한 NVIDIA의 해법입니다. 이는 딥 러닝 기반 추천 시스템에 중점을 두고 시작됐습니다. 제가 큰 관심을 갖는 분야이기도 합니다.” 단순하게 시작하라는 업계 추천 시스템 전문가들의 핵심 조언에 따라 Merlin의 기능은 고객의 프로젝트 수명주기 전반에서 고객의 필요 사항을 아우릅니다. "Merlin은 간단한 모델로 시작하여 시간이 지남에 따라 더 복잡하게 구축할 수 있도록 프레임워크를 제공합니다. 이 분야에는 빈틈이 있습니다. 현재 기존 머신 러닝에서 딥 러닝으로 전환하기가 굉장히 어렵습니다. 많은 회사들이 이 단계에서 고전하고 있으며, 딥 러닝 기반 작업을 수행하려면 완전히 다른 파이프라인을 구축해야 하는데 그것은 아주 복잡합니다.” 또한 비용도 많이 들며, Monica Rogati 등 전문가들이 경고하듯이 프로젝트 로드맵에서 이렇게 불연속적인 지점은 위험으로 이어질 수 있습니다.

Even은 다음과 같이 말했습니다. “저희는 팀이 좋은 추천 시스템을 구축할 때 따르는 문제를 해결하려는 것입니다.” 이를 어떻게 쉽게 만들 수 있을까요? 어떻게 해야 최종 솔루션뿐 아니라 파이프의 각 단계에서 솔루션을 제공할 수 있을까요? "추천 시스템을 개발하고 배포하는 것은 오늘날 매우 복잡한 작업입니다. 머신 러닝도 복잡하지만, 추천 시스템은 더 복잡합니다. 하지만 이제 상황이 변하고 있습니다. 저희는 사용과 구축이 쉽고 성능이 뛰어난 프레임워크인 Merlin에 총력을 기울이고 있습니다.”

이 접근법의 한 가지 중요한 측면은 모든 고객이 저마다 로드맵을 가져야 한다는 것입니다. 오늘날의 요구 사항은 계속해서 변화하니까요.

"HugeCTR 팀의 작업은 NVIDIA가 보다 수준 높은 추천 시스템에 주력하고 있음을 보여주는 대표적 사례입니다. 현재 이 팀이 작업 중인 프로젝트는 100테라바이트 모델입니다. 현재 그 어떤 NVIDIA 고객보다도 큰 규모지만, 고객들도 머지않아 이 수준에 도달할 것입니다. 그리고 그 규모에 도달하면 아주 스마트하게 생각해야 합니다. 이 팀은 어떤 팀이 오픈 소스에서 작업하고 발행하고 있는지 보고 이를 제품에 통합하고 있으며, 다른 팀에 지침도 제공합니다. 일부 팀은 이 기술을 자신들의 스택에 통합하여 채택합니다. 이 수준에서는 많은 작업이 큰 규모에서 이뤄지며 여기에 주력하는데, 그럴 만합니다. 아주 많은 비즈니스가 있으니까요. 추천 시스템으로 작업하는 기업의 규모는 굉장히 다양합니다. 그리고 큰 규모의 기업에서 요구되는 컴퓨팅 수준은 작은 기업과 비교가 되지 않습니다.”

> 중요한 장애물의 해결

복잡한 요구를 처리하는 머신 러닝 모델을 구축하기는 어렵습니다. 언급했듯이 모델을 구축하는 것만으로는 충분하지 않습니다. 지금까지는 일반적인 머신 러닝 워크플로우에서 주목을 덜 받는 부분이 있었습니다. 오픈 소스 에코시스템이 번성하려면 이 빈틈을 메우는 것이 매우 중요합니다. “이 모델을 어떻게 프로덕션 환경에 도입할 수 있을까요? 이것은 많은 회사에게 아주 큰 과제입니다. 그리고 쉽지 않은 과제입니다. 딥 러닝 측면에서는 특히 그렇습니다. 기초적인 협업 필터링조차 뭔가를 생산 환경에 구현하고 적절하게 모니터링하여 궤도에서 벗어나지 않게 해야 하고, 추적해야 하는 모든 것들을 제대로 추적하고 있는지 확인해야 합니다. 이것은 대단히 복잡합니다. 그래서 [Merlin] 팀의 이차적인 작업은 이를 생산 환경에 쉽게 구축하도록 만드는 것입니다.” 추천 시스템 활용의 이러한 측면에서 위험을 제거하는 것 또한 Merlin 팀의 과제 중 하나입니다.

> 추천 프로젝트 시작하기

추천 시스템 여정을 시작하는 사람들에게 Even은 다음과 같이 조언합니다. "핵심은 단순하게 시작해서 반복 작업하는 것입니다. 데이터 사이언티스트들은 가장 훌륭한 최신 딥 러닝 모델을 구축하고자 합니다. 보다 단순하고 간단한 것부터 시작해서 장래성이 있어 보이는지 확인하세요. 제 경험상 좋은 데이터는 항상 좋은 모델보다 낫습니다. 데이터 사이언티스트는 종종 험난한 경험 끝에 그런 교훈을 얻습니다. 저는 이것이 아주 보편적인 사실이라고 생각합니다. 따라서 데이터를 잘 정리하고 관리한 다음 특징 탐색을 하는 것은 최신 모델을 사용하고 보자는 관점으로 시작하는 것보다 더 나은 결과로 이어질 가능성이 높습니다.

Even은 경력 초기에 인기 온라인 데이팅 시스템에 사용될 추천 시스템을 구축했습니다. Even은 실무자에게 수집되는 데이터, 측정되는 메트릭스, 구축 환경, 사용 사례에 대한 목표 함수 등을 파악할 수 있는 프로세스를 미리 마련할 것을 권장합니다. 모델은 무엇인지, 어떻게 반복 작업할 것인지, 결과물은 어떤 유형의 모델이 될 것인지 이해하는 것입니다. 그런 다음 생산을 담당하는 팀과 긴밀히 협력하여 프로세스, 제약 조건 등을 파악합니다.

요약

지금까지 살펴본 인터뷰는 소매, 뉴스, 소셜 미디어, 게이밍 등 다양한 비즈니스 분야의 다양한 추천 시스템 활용 간에 많은 공통점이 있음을 시사합니다.

무엇보다도 오픈 소스는 업계의 추천 시스템을 위한 필수 구성 요소가 되었습니다. 오픈 소스 에코시스템 내에서 도구의 상호 운용성은 특히 전체 수명 주기를 고려할 때 프로젝트의 위험을 낮추는 데 필수적입니다. 이러한 도구는 반드시 실험과 탐구를 뒷받침하면서 아직 주목받지 않은 신기술도 받아들일 수 있을 정도로 유연해야 합니다.

또 한 가지는 데이터의 문제입니다. '빅 데이터' 개념은 2000년대 중반에 나타났습니다. 그 후로 생산 환경에서 머신 러닝은 더 나은 품질의 데이터를 더 많이 확보할수록 모델의 효율성이 높아진다는 사실은 보편적으로 인정되고 있습니다. 추천 시스템의 경우도 마찬가지입니다. 더 나은 알고리즘만으로 모델을 개선할 수는 없습니다. 그 대신 좋은 학습 데이터, 추론을 위한 좋은 고객 데이터, 메트릭스 평가를 위한 좋은 피드백과 기구 등이 필요합니다. 효율적인 데이터 준비 작업도 매우 중요한 요구사항입니다. 추천 시스템 사용 사례가 성공할수록 데이터 품질 또한 증가할 것입니다. 또한 성공을 위한 측정 기준을 세분화할 때 데이터를 통해 지속적으로 학습하는 것도 중요합니다. 프로덕션 환경의 추천 시스템을 철저하게 이해할 수 있도록 데이터 사이언스 직원을 '메트릭스 엔지니어'로 재편성하는 것도 도움이 될 수 있습니다.

전문가 인터뷰에서 강조된 주요 사항을 요약하면 다음과 같습니다.

요약 #1: 단순하게 시작하기

많은 전문가들이 추천 시스템 시작에 대해 일관되게 조언하는 내용은 단순하게 시작하라는 것입니다. 무엇이 필요한지 생각하세요. 업계 보고서에서 추세가 되는 복잡한 머신 러닝 모델을 구현하기 위해 서두르는 대신 간단한 규칙 기반 기준을 먼저 따라해 보세요. 프로세스를 따라 테스트, 측정, 반복하세요. 그러면 솔루션이 보다 적은 위험으로 단순한 것에서 복잡한 것으로 발전할 수 있습니다.

Pete Warden은 JetPac에서 시작되어 Google로 이어진 관련 관행에 대한 글을 남겼습니다. '오즈의 마법사 방법'이라 불리는 이 관행에서 팀은 우선 실시간 대상 그룹 앞에서 대화형으로 수행된 제안된 제품의 머신 러닝 구성 요소를 모의 구성합니다.

한편, 잘 알려지고 검증된 기술에서 시작하세요. 통합의 용이성을 우선시하고 결과에 들어가는 수고를 최대한 줄이세요.

요약 #2: 목표 함수 파악

성공을 위한 메트릭스를 결정하고 주어진 사용 사례에 대한 목표 함수를 철저히 이해합니다. 가장 많이 반복되고 언급된 조언입니다.

요약 #3: 비즈니스 모델로 확장

Netflix Prize (2006~2009년) 를 돌이켜 보면, 이 추천 시스템 대회는 포인트 솔루션으로 형성됐습니다. 즉 추천 시스템은 순위가 매겨진 항목 목록을 제공하는 것으로 여겨졌습니다. 입력과 출력이 있고, 그 사이에 '블랙 박스' 솔루션이 있었습니다.

우리 모두는 2009년 이래로 프로덕션 환경의 추천 시스템이 비즈니스 모델을 어떻게 바꾸었는지 직접 봐 왔죠. 예를 들어 식료품점은 일반적으로 마진이 적은 것으로 간주됩니다. 식료품 배달 비즈니스 모델을 구축하려는 초기 시도는 어려움을 겪었습니다. 추천 시스템을 도입함으로써 고객과의 접점뿐 아니라 비즈니스 프로세스의 보다 깊은 부분에도 머신 러닝을 적용할 수 있게 됐습니다.

이제 추천 시스템은 사용자 환경 전반에 걸쳐 훨씬 더 광범위한 범위에 개인화를 적용합니다. 추천 시스템은 긴급 사태에 대비한 계획을 수립하고 비즈니스의 다른 측면을 최적화하도록 돕습니다. The New York Times의 경우 머신 러닝 솔루션이 독자의 신뢰와 충성도를 구축하는 데 최적화된 추천 시스템을 내세워 비즈니스의 경제력을 뒷받침합니다. 이는 편집부의 판단을 대체하는 것이 아니라 확장하는 방식으로 이루어집니다.

불과 몇 년 만에 머신 러닝 활용이 의문의 대상에서 핵심 비즈니스 전략으로 바뀌었습니다. 이 기간 동안 하드웨어의 진화는 훨씬 복잡하고 강력한 머신 러닝 기술의 발전과 함께 비즈니스 문화의 인식을 바꿨습니다.

요약 #4: 빠른 추론 및 하드웨어 가속화

역사적으로 오픈 소스 프로젝트와 벤더 제품은 ETL, 데이터 준비 등 일반적인 머신 러닝 워크플로우의 초기 단계에 보다 초점을 맞춥니다. 고객 경험에 가장 가까운 모델 서비스 및 실시간 추론 등 후반 단계에서는 최종 사용 애플리케이션의 일부인 임시 코드로 격하되는 경향이 있습니다. 생산에서 많은 추천 시스템 사용 사례는 보통 100밀리초 미만의 빠른 추론을 필요로 합니다.

추천 시스템에서 활용할 수 있는 많은 옵션과 흥미로운 기술이 있지만, 추론에 필요한 시간은 게이트 함수이다. 품질 높은 추천이 주어진 시간 간격 내에 생산될 수 없다면 보다 간단한 기술을 대신 사용해야 합니다. 예를 들어 정교하고 리소스가 잘 갖춰진 관행에서도 최근 2년 전 정도까지만 해도 머신 러닝 모델 대부분은 상대적으로 단순한 선형 모델이었습니다. 그러나 최근에는 딥 러닝 모델이 도입되었습니다. 획기적인 변화 요인은 빠른 추론을 가능케 하는 하드웨어였습니다. 이러한 하드웨어는 보다 정교한 AI 채택을 실현하며, 이는 추천 시스템이 비즈니스 모델 및 운영의 보다 깊은 부분으로 확장된다는 앞의 요약과 연결됩니다.

머신 러닝 분야가 계속 발전함에 따라 앞으로 분명 보다 정교한 접근법이 계속 생겨날 것입니다. 예를 들어 GAN, 자기 지도 학습, 강화 학습은 아직 여전히 연구 단계에 있지만 각각 추천 시스템에 대한 흥미로운 응용 사례를 제시합니다. 이런 구체적인 예시는 더 많은 컴퓨팅 리소스를 필요로 합니다.

질문: 고급 기술은 밀리초 내에 실행하기에 충분할 정도로 성능이 뛰어날까요? 그러한 제약으로 인해 프로덕션 환경에서 추천 시스템이 최신 혁신을 채택하기가 어려울까요? 그 대신 방대한 메모리 공간 및 하드웨어 가속을 제공하는 조정된 클러스터는 보다 고급인 AI 애플리케이션을 실현하는 요소가 될 것입니다.

요약 #5: 주요 구성 요소

업계에서 추천 시스템 관행과 관련하여 주목받는 주제로는 특징 스토어 사용 (수년 전 소개됨), 그래프 신경망 사용, 모델 증류 등이 있습니다.

요약 #6: 로드맵 계획 수립

시스템 업그레이드를 미리 계획합니다. 보다 복잡한 딥 러닝 모델로 대체되는 단순 선형 모델과 같은 구성 요소의 측면을 고려하면 운영 환경에 변화가 필요할 수 있습니다. 아주 다른 의미에서 추천 시스템은 비즈니스의 경제적 활력을 지원하므로 그러한 활용은 성장할 것으로 보입니다. 데이터 전송률은 증가할 것입니다. 시간이 지남에 따라 추천 시스템은 비즈니스 프로세스로 더 깊이 확장될 것입니다. 또는 다른 비즈니스 부서에 부가적인 추천 시스템 사용 사례가 필요할 수 있습니다.

시스템 업그레이드 및 증가하는 사용 사례의 로드맵을 계획하세요. 가능한 현재의 확장 요구 사항뿐 아니라 성공하는 비즈니스가 몇 년 앞서 필요로 할 만한 것을 고려하면서 확장을 위한 로드맵을 제공하는 기술을 활용해야 합니다.

전 세계 수십억 명의 사람들이 온라인에 참여하고 브라우징, 쇼핑, 학습, 친구와 소통 등으로 소비하는 순간들에서 각 순간, 각 이벤트, 각 세션은 정보를 바탕으로 추천 시스템이 각 개인을 위해 보다 쉽고, 빠르고, 보다 개인화된 결정을 내리도록 지원할 기회가 됩니다. 이는 수십억 명의 사람들이 온라인에서 수조 개의 물건들과 상호작용한다는 점을 시사합니다.

자세히 알아보기

NVIDIA Merlin에 대해 보다 자세히 알아보려면 다음 웹 사이트를 방문하세요. developer.nvidia.com/nvidia-merlin

© 2021 NVIDIA Corporation. All rights reserved. NVIDIA, NVIDIA 로고, Merlin 및 [기타 NVIDIA 제품 및 기술]은 미국 및 다른 나라에서 NVIDIA Corporation의 상표 및/또는 등록 상표입니다. 다른 회사 및 제품 이름은 연관된 각 회사의 상표일 수 있습니다. 기타 모든 상표는 해당 소유자의 자산입니다.
SEP21

선임 연구원 및 전문가와의 심층 인터뷰

NVIDIA

본 보고서의 전반부에서는 트렌드와 각 분야의 선임 연구원들 및 전문가들과의 인터뷰를 요약했습니다. 이 섹션에서는 심층 인터뷰 일부를 다뤘습니다. 인터뷰 내용은 가독성을 위해 인터뷰 대상자의 승인을 받아 편집했습니다.

Q: 추천 시스템을 구축하는 데 사용한 데이터의 크기는 어느 정도입니까?

추천 시스템을 아주 폭넓게 정의한다면 제가 사용한 데이터의 범위는 제로 데이터 포인트부터 수백만 명의 사용자와 항목 간의 일치까지 다양했습니다. Monica는 "이 분야의 최상위 기술은 굉장히 많이 주목을 받습니다. 잔존 문제를 해결할 수 있고 마법처럼 (또는 관점에 따라서는 불쾌하게) 느껴질 정도로 굉장히 개인화된 추천을 제공하기 때문입니다." 또한 데이터 사이언티스트가 이끌리는 흥미로운 알고리즘은 보다 방대하고 다양한 데이터에서 빛을 발합니다.

Q: 작업을 지원하기 위한 적절한 기술과 프레임워크를 어떻게 선택하십니까?

저는 매우 실용적으로 기술과 프레임워크를 추천합니다. 문제 해결 가능성 외에도 통합의 용이성, 성숙도, 선택권 보존 등 고려해야 할 몇 가지 기준이 있습니다.

가장 중요한 기준은 회사의 현재 워크플로우와의 통합 용이성과 결과 대비 투입 리소스의 비율을 최소화하는 것입니다. 한 회사가 LinkedIn의 저희 팀에 특정 기술을 판매하려고 한 적이 있었는데, 저희는 대규모 문제를 겪고 있어서 이 기술을 굉장히 사용해 보고 싶었지만 상당한 시간과 노력을 할애하지 않고서는 이를 테스트할 방법이 없었습니다. 여러 차례 검증되고 잘 알려진 기술을 활용할 때는 그런 위험을 감수할 수 있지만 새로운 기술은 그렇지 않으니까요. 데이터 사이언스 분야에서 "지루한 기술을 사용하라"는 말은 "야채를 많이 먹으라"는 말과 같습니다. 모두가 이론적으로는 좋은 생각임을 알고 있지만, 훌륭하고 재미있는 최신 기술이 엔지니어와 데이터 사이언티스트를 유혹합니다. '지루한' (성숙한) 기술은 대규모 사례에서는 특히 더 중요합니다. 핵심 경로에서 벗어난 작은 문제에는 언제든 최신 도구를 시도해 볼 여지가 있습니다. 트레이드오프를 살펴보고 이용하는 것은 단순히 권장 시스템 문제가 아닙니다. 새로운 도구를 사용해 보는 것을 혁신을 위한 일종의 투자라고 생각해 볼 수도 있죠. 어쩌면 데이터 팀의 사기를 북돋고, 더 많은 인재를 채용하고, 보유율을 향상시키는 일로도 말입니다.

"지난 몇 년 동안 데이터 도구와 프레임워크가 폭발적으로 증가했습니다. 대다수 도구가 특정 환경에서 입증되고 신뢰할 수 있을 때까지는 워크플로우에 큰 변화 없이도 쉽게 통합될 수 있게 됐다는 점은 좋은 일입니다. 이런 도구를 만드는 분들에게 제가 항상 묻는 질문은 이렇습니다. '여러분의 트로이 목마는 무엇인가요? 채택을 받고 신뢰를 얻기 위해 실무자의 노력을 거의 또는 전혀 들이지 않고도 해결 가능한 *한 가지* 문제는 무엇인가요?

마지막으로, 선택지를 보존하는 것은 프레임워크나 기술을 채택하는 초기에 중요합니다. 채택한 기술이 가장 적합하지 않거나 요구사항이 변경됐다는 것이 밝혀지면 어떻게 할까요? 이는 단지 특정 기술에 종속되는 것을 말하는 게 아니라, 프로세스 초기에 기술의 위험 요소를 제거하고 예비 계획을 수립하는 것을 말합니다. 최고의 도구는 마케팅의 잠재적 탈출구를 포함하며 쉬운 통합 지점뿐 아니라 데이터 내보내기 도구 또한 제공합니다.

Q: 오픈 소스와의 상호 운용성은 얼마나 중요합니까?

매우 중요합니다. 특히 적절한 기술과 프레임워크를 선택하는 맥락에서는 더욱 그렇습니다. 다양한 오픈 소스 기술에 플러그 앤 플레이할 수 있을 때 '쉬운 통합'은 자연스럽게 실현됩니다. 마찬가지로 모듈형 기술에서 퍼즐 조각을 대체해도 나머지 시스템에 거의 영향을 미치지 않는다는 사실을 아는 것은 위에서 제시한 '선택지 보존'이라는 기준을 충족하는 데 큰 도움이 됩니다.

MONICA ROGATI

AI 및 데이터 사이언스 고문

LinkedIn "알 수도 있는 사람" 기능의 머신 러닝 모델 최초 개발자

Monica Rogati는 다양한 업계에서 5~8000명의 직원 보유한 회사들에게 기술 AI 및 데이터 사이언스 관련 조언을 제공합니다. 독립 고문 활동 이전에는 Jawbone의 데이터 부사장으로 최고의 데이터 과학자와 엔지니어 팀을 구성하고 이끌었습니다. Jawbone 이전에 Monica는 LinkedIn 데이터 사이언스 팀의 초기 멤버였으며 LinkedIn 회원을 위한 일자리 매칭, 알 수도 있는 사람 찾기, 전문가 그룹 추천하기 등 LinkedIn의 핵심 데이터 제품을 개발 및 개선했습니다. Monica의 전문 분야는 응용 머신 러닝 (Carnegie Mellon 컴퓨터공학 박사), NLP, 웨어러블 기기 등입니다. Monica가 개발한 AI 욕구 단계 다이어그램은 데이터 사이언티스트를 오해로부터 보호합니다.

Q: 모델을 어떻게 평가하십니까?

이 지점에 도달했다면 훈련 데이터를 윤리적인 방식으로 획득하고 조합하여 문제를 조사한 상태일 것이며, 주어진 목표 기능을 극대화했을 때 시스템은 유해하지 않을 것입니다. 이는 생각보다 훨씬 더 어려운 문제입니다. '좋은 의도'는 만족스러운 솔루션과 거리가 있습니다. 전문가가 인정하는 모범 사례는 지난 몇 년 동안 발전해 왔기 때문에 추가 연구가 필요할 것입니다.

모델 평가에 대한 질문으로 돌아가자면, 저는 기준선, 넉넉한 오프라인 필터, 느린 증가 속도를 사용한 A/B 테스트라는 세 가지 단계를 권장합니다.

첫 단계는 데이터를 분석한 다음 합리적이면서 말도 안 되게 간단한 규칙 기반 기준선을 만드는 것입니다. 추천에서 이는 단지 '지난 3개월 동안 가장 인기 있었던 상품'일 수도 있습니다. 아니면 추가 정보가 사용 가능한 경우 몇 가지 속성을 규칙에 추가할 수 있습니다. "지난 3개월 동안 사용자의 국가/지역에서 가장 인기 있었던 상품" 같은 식입니다. 긍정적인 예시가 비교적 적은 분류 문제의 경우, 이는 '항상 NO라고 말하라'가 될 수 있습니다 (이는 비교적 드문 이벤트에서 99% 정확도를 주장하는 것이 기술적으로는 참일지라도 오해의 여지가 있기 때문입니다).

이미 모델이 있는데 이 단계를 건너뛰었다면, 그래도 한번 시도해 보세요. 간단한 기준선은 디버깅이 쉽고, 견고하고, 빠릅니다. 현재 모델만큼 성능이 좋다면 이를 선호합니다.

'넉넉한 오프라인 필터'는 기존의 오프라인 모델 평가를 살짝 변경한 것입니다. 여기서 목표는 모델이 엄청난 혼란이나 예기치 못한 피해를 일으키는 실패작이 아니고 아주 넓은 의미에서 '유망한지'를 확인하는 것입니다. 꼭 현재의 시스템을 극복하는 것이 목표는 *아닙니다*. 그러나 위의 기준선에서 시작하는 경우에는 그래야 합니다.

마지막으로, 진정한 평가는 아주 느리게 진행되는 A/B 테스트입니다. 운영 문제, 누락된 기능, 예기치 않은 피해, 기타 일반적으로 오프라인 또는 소규모에서는 나타나지 않을 수도 있는 일반적인 문제에 주목합니다. 모델이 본격적으로 실행된 후, 특히 추천 시스템의 경우 이상치 효과가 사라질 수 있도록 'Burn in' 기간이 필요한 경우가 많습니다.

Q: 추천 시스템에서 해결하기 어려운 문제가 있었다면 무엇입니까?

가장 큰 과제는 올바른 메트릭스 또는 목표 함수를 설계하는 것입니다. 데이터 사이언티스트는 결국 메트릭스 엔지니어가 될 것입니다. 이는 자동화하기가 불가능하지는 않더라도 굉장히 어려운 데이터 작업으로 꼽힙니다.

Q: 팀장이 이제 막 시작해서 현재 추천 시스템 구축, 배포, 최적화를 평가 중이라면 어떤 조언을 하시겠습니까?

최대화하려는 메트릭 (목표 함수) 에 대해 신중하게 생각하세요. 종종 *실제로* 달성하려는 것의 프록시 (Proxy) 를 측정하는 것이 더 쉽습니다. 당신의 시스템에 영향을 준 후에도 여전히 좋은 프록시가 될까요? 인센티브 일치로 인해 자발적 또는 우발적으로 조절할 방법은 있는가? 문제는 얼마나 어려운가? 성능에서 고려하는 상한은 어느 정도인가? 예를 들면 학습 데이터의 채점자 간 합의는 어느 정도인가? 이런 문제를 고려해야 합니다.

IBM Research에서 인턴으로 일하면서 Salim Roukos에게 배운 것이 있습니다. 머신 러닝 시스템으로 작업을 시작하기 전에 1시간 동안 데이터 레이블링과 알고리즘의 작업을 수행하라는 것입니다. 알고리즘 그 자체가 되세요. 수많은 데이터 품질 문제를 발견하거나, 질문이 잘못 정의됐을 때의 상황들을 식별하거나, 언제 추천을 하지 말아야 하는지 등을 알 수 있습니다.

XIANGTING KONG
전문가 엔지니어
Tencent

Q: Tencent에서 어떤 직책을 맡고 계신지요?

저는 Tencent의 전문 엔지니어이며 광고 추천 시스템의 설계 및 개발을 담당하고 있습니다. 또한 Tencent 광고 및 딥 러닝 플랫폼을 이끌고 있습니다. 저희 플랫폼은 광고부터 핀테크, 네트워크 데이터 마이닝 등 다양한 비즈니스 시나리오에서 머신 러닝 모델 최적화, 훈련, 추론을 지원합니다.

Q: Tencent에서 어떤 작업을 하고 계십니까?

저희 팀은 주로 머신 러닝 플랫폼을 개발하며 특징 엔지니어링, 모델 훈련 및 온라인 추론을 담당합니다. 0에서 1까지 GPU를 기반으로 광고 추천을 하기 위한 차세대 고성능 분산 훈련 시스템을 구현하려 노력하고 있습니다.

Q: Tencent의 전반적인 비즈니스와 팀의 추천 시스템 작업과는 어떤 관련이 있습니까?

Tencent의 광고 추천 훈련 플랫폼은 전체 Tencent 트래픽 비즈니스를 아우릅니다. Tencent 광고 추천은 WeChat, Moments, QQ, Tencent Games, Tencent Video, Tencent News 등과 같은 서비스에서 널리 사용됩니다. Tencent의 광고 매출은 수십억에 달합니다. 광고 추천의 정확성은 광고 수익을 높이는 데 도움이 됩니다.

Q: 팀은 비교적 새로운 팀입니까? Tencent는 왜 추천 시스템에 투자하기로 결정했나요?

저희 팀은 수년에 걸쳐 설립되었습니다. 광고 사업은 Tencent 내에서 비교적 중요한 사업이며 추천 시스템은 전체 광고 수익을 높이는 데 사용됩니다.

Q: 팀에서 어떤 종류의 추천 시스템에 초점을 맞추고 있습니까?

저희 팀은 광고 훈련 플랫폼의 최적화를 담당하는 광고 추천 시스템에 집중하고 있습니다. Tencent 광고 추천 시스템은 오프라인 특징 엔지니어링, 교육 플랫폼, 온라인 추론 시스템, 온라인 특징 엔지니어링 및 플레이 플랫폼 등으로 구성됩니다. 광고 추천은 점진적인 필터링 과정입니다. 정렬 단계에는 리콜, 사전 순위 지정 및 순위 지정이 포함됩니다. 각 단계마다 요구 사항이 다릅니다. 모델의 신속한 조사와 반복은 훈련 성능을 위한 더 높은 요구 사양을 요구합니다.

Q: 팀에서는 어떻게 훈련을 실시합니까?

저희는 1~2주마다 서로 기술을 공유합니다.

Q: 팀에서 추천 시스템을 어떻게 평가하시는지요? 미세 조정은 어떻습니까?

추천 시스템을 통해 알고리즘 전략을 최적화하고, 샘플과 특징을 더 추가하고, 수익 증가를 유도할 수 있는지 여부를 평가합니다. 샘플 특징을 추가하여 더 많은 샘플 데이터를 훈련시키면 광고 추천의 정확도를 높일 수 있습니다. 그러나 이는 훈련 시간을 늘리고 모델의 업데이트 빈도에 영향을 미칩니다. 모델 업데이트가 궤도를 벗어나지 않도록 하려면 모델의 훈련 성능을 지속적으로 향상시켜야 합니다. 훈련 모델 성능이 향상되면 모델의 정확도를 향상시키기 위해 더 많은 데이터를 학습하여 광고 수익을 높일 수 있습니다.

Q: 추천 시스템은 어떻게 최적화하십니까? 예를 들어 Tencent는 최적화를 포함하기 위해 HugeCTR을 사용하고 있는데요. 이것이 워크플로를 최적화하는 데 어떻게 도움이 되었습니까?

HugeCTR은 추천 훈련 프레임워크로서 모델 훈련의 업데이트 빈도를 높이기 위해 광고 추천 훈련 시스템에 통합되어 있으며, 더 많은 샘플을 학습하여 온라인 효과를 개선할 수 있습니다.

Q: 작업을 지원할 적절한 기술, 패키지, 방법, 프레임워크는 어떻게 선택하십니까?

선택하는 기술이나 프레임워크가 커뮤니티 생태계와 호환되어야 더 나은 후속 업그레이드를 수행할 수 있습니다.

Q: 모델 확장은 어떻게 해결하십니까?

더 큰 모델을 사용하면 더 많은 특징을 학습하는 데 도움이 되며, 따라서 모델의 정확도가 향상됩니다.

Q: 팀의 최근 성공 사례는 무엇입니까?

훈련 프레임워크에서 데이터 병렬 분산 솔루션 (Data-parallel distributed solution) 을 개발했습니다.

Q: 최근에 추천 시스템 워크플로우에 특정 방법들을 통합하셨나요?

최근 CSR (압축 희소 행렬) 파이프라인을 광고 추천 훈련 시스템에 통합했습니다. 훈련용 GPU에서 데이터를 직접 읽을 수 있도록 CSR 유형 훈련 데이터를 생성합니다. 데이터 처리 파이프라인 최적화를 통해 CPU 로드가 크게 줄었고 GPU 활용은 대폭 향상됐습니다.

Q: 팀장이 이제 막 시작해서 현재 추천 시스템 구축, 배포, 최적화를 평가 중이라면, 추천 시스템 워크플로우를 가속 및 간소화하는 작업에 대해 어떤 조언을 주시겠습니까?

성숙한 기술 프레임워크를 선택하고 커뮤니티 생태계와 호환을 유지하여 후속 시스템 업그레이드를 용이하게 하세요.

CHRIS WIGGINS

수석 데이터 사이언티스트

The New York Times

Q: The New York Times에서 현재 어떤 직책을 맡고 계십니까?

수석 데이터 사이언티스트로서 뉴스룸 및 비즈니스 문제를 위한 머신 러닝 솔루션을 개발 및 배포하는 팀을 이끌고 있습니다.

Q: The New York Times의 전반적인 비즈니스와 팀의 추천 시스템 작업과는 어떤 관련이 있습니까?

The Times는 2025년까지 1,000만 명의 유료 구독자를 달성하는 것을 목표로 하고 있습니다. 구독자들은 The Times 편집부의 판단에 관심이 많습니다. 따라서 추천은 신뢰와 충성도에 맞춰 최적화하여 편집부의 판단을 대체하지 않고 강화하는 방향으로 일하도록 만드는 데 그 의미가 있습니다. 현재 추천은 "더 많은 내용", "편집자의 선택", "더 스마트한 생활", 국제 홈 페이지, "가장 인기 있는 기사" 등 다양한 뉴스 표면에 표시됩니다. 또한 요리 앱, 모바일 앱의 "당신을 위한 뉴스" 탭 및 "당신의 주간 에디션" 뉴스레터에서도 추천을 활용합니다.

Q: 팀은 비교적 새로운 팀입니까? The New York Times는 왜 추천 시스템에 투자하기로 결정했나요?

그렇기도 하고 그렇지 않기도 합니다. 제가 NYT에 있는 동안 개인화는 수많은 '역동적 팀 재구성'을 통해 계속 발전했습니다. 그래서 지속되는 계보는 있지만 라인업과 역할은 바뀌었습니다. 마치 영국 록 밴드 포스터와 같죠. 차이가 있다면 Jimmy Page는 Yardbirds를 탈퇴해서 새로운 즐거움을 쫓았지만, 개인화는 인프라 역량과 제품 목표를 원동력으로 계속 발전한다는 점입니다. 6월에는 데이터 사이언스 부서를 재조정해서 홈페이지의 개인화를 이끌었던 Anna Coenen 박사가 이제 알고리즘 추천 팀을 담당하고 있습니다. 이전에 이 팀을 맡았던 Anne Bauer 박사는 이제 현장 마케팅을 위한 머신 러닝 팀을 이끌고 있습니다. 추천 시스템은 제가 입사한 2013년보다 한참 이전부터 디지털 전략의 일부였습니다. 현재는 사용자 경험과 비즈니스 목표의 일부로 간주됩니다.

Q: 작업을 지원할 적절한 기술, 패키지, 방법, 프레임워크는 어떻게 선택하십니까?

전략은 Kreps 법칙입니다. 저는 이 법칙을 키보드 단축키로 지정해 뒀습니다.

전술은 제품과 뉴스룸 파트너의 목적에 달려 있습니다. 뉴스룸 파트너는 'x' (맥락), 'a' (어떤 활동, 기사, 자산을 추천할 것인지), 'y' (어떤 결과를 최적화하고자 하는지) 를 깊이 생각합니다.

Q: 모델 확장은 어떻게 해결하십니까?

저희 인프라는 Google 클라우드 서비스를 통해 배포됩니다. 데이터 수집은 2018 Open 블로그 게시물에서 설명한 대로 전용 이벤트 추적 서비스를 통해 이뤄집니다. 추천 API에 대한 요청은 추천을 지원하는 동시에 Samizdat이라는 내부 서비스를 통해 사용자 개인 정보를 보호합니다. 이 서비스에 대해서는 2021 Open 블로그 게시물에 설명이 되어 있습니다.

Q: 최근의 성공 사례나 성공적으로 해결된 과제는 무엇인가요?

한 가지 성공은 예측적 방법에서 규범적 방법으로 전환한 것이었습니다. 예를 들면 다양한 상황별 밴디트 (2019 Open 블로그 게시물 참조) 를 콘텐츠의 서로 다른 표면 및 풀에 대해 배포할 수 있습니다. 구조가 없는 피드는 하나도 없기 때문에 확장에서 문제는 새 알고리즘과 새 표면에서 반복 작업하는 것이었습니다. 여기서는 뉴스의 '유통기한' 등 데이터에 대한 생각과 최적화되는 대상에 대한 생각이 큰 차이로 이어질 수 있습니다. 또 한 가지 성공은 Cooking 앱에 추천을 구축한 것입니다. '독립형' 제품으로서 아주 성공적인 제품이었으며 갈수록 NYT 인프라에 통합되며 다수의 새로운 제품 개발 실험을 가능케 하고 있습니다. 이 콘텐츠는 '뉴스'와 매우 다르기 때문에 저희가 편집부의 비전을 포착하고 강화할 수 있도록 데이터 사이언스와 제품 간에 창의적인 파트너십이 많이 이뤄졌습니다.

Q: 최근에 워크플로 통합을 고려하고 있는 특정 기술이나 방법에 대한 추천 시스템 논문이 있나요? 아니면 최근에 특정 방법을 통합하셨는지요?

그렇기는 하지만 아직 공개하지는 않고 있습니다. open.nytimes.com을 주목해 주세요.

Q: 팀장이 이제 막 시작해서 현재 추천 시스템 구축, 배포, 최적화를 평가 중이라면, 추천 시스템 워크플로우를 가속 및 간소화하는 작업에 대해 어떤 조언을 주시겠습니까?

저는 Kreps 법칙이나
Monica의 인포그래픽 등을 활용합니다. NYT의 성공 비결은 데이터 사이언티스트와 소프트웨어 엔지니어가 나란히 앉아 함께 개발하고, Python (데이터 사이언티스트) 과 Go (소프트웨어 엔지니어링 파트너) 를 혼합하여 코딩함으로써 '통계와 SLA' 양쪽에서 성과를 내는 방법을 구축한 것입니다.

JUN HUANG

Meituan 수석 기술 전문가

Q: Meituan은 어떤 기업입니까? 얼마나 많은 사용자와 판매자가 Meituan을 사용하고 있는지요? Meituan 기술 플랫폼에서 얼마나 많은 거래가 발생합니까?

Meituan의 사업 범위는 매우 광범위하여, 중국 전역 2,800개 이상의 도시에서 케이터링, 온디맨드 배송, 자동차 운송, 자전거 공유, 호텔 및 여행 예약, 영화 발권 및 기타 엔터테인먼트 및 라이프스타일 서비스 등 200가지 이상의 서비스 카테고리를 운영합니다. Meituan은 연간 6억 3000만 명의 활성 사용자와 770만 개의 활성 비즈니스를 보유하고 있습니다. 각 사용자는 연간 32.8건의 거래를 기록합니다.

저희의 사명은 사람들이 더 잘 먹고 잘 살도록 돕는 것입니다.

Q: Meituan에서 맡은 직책은 무엇입니까?

저는 Meituan 머신 러닝 플랫폼의 훈련 프레임워크 팀을 담당하는 Meituan의 수석 기술 전문가입니다. 추천 시스템, NLP, CV, ASR, 자율주행 등 Meituan의 많은 딥 러닝 분야를 다룹니다.

Q: 팀은 Meituan에서 어떤 작업을 하나요?

저희 팀은 안정적인 고성능 분산 딥 러닝 훈련 프레임워크를 개발했습니다. 저희 팀의 시스템은 대규모 CPU/GPU 클러스터에서 배포가 가능하며 실패 예비 조치 및 자동 스케일링을 지원합니다. 추천 시스템 시나리오에서는 최대 1000억 개의 희소 파라미터와 최대 1000억 개의 샘플로 분산 훈련을 지원하며, 온라인 학습도 지원합니다. NLP 시나리오에서는 수백 개의 GPU에서 100억 개의 매개 변수를 갖춘 분산 교육을 지원합니다. 최근 NVIDIA A100을 기반으로 차세대 추천 훈련 시스템을 설계하여 훈련 효율성과 모델 복잡성을 크게 향상시켰습니다.

Q: Meituan의 전반적인 비즈니스와 팀의 추천 시스템 작업과는 어떤 관련이 있습니까?

저희 훈련 프레임워크는 Meituan의 전반적인 트래픽 비즈니스인 검색, 추천 및 광고 시나리오를 위한 모델 훈련을 다룹니다. 저희는 Meituan이 성장과 실현을 가속화할 수 있도록 도왔습니다.

Q: 팀은 비교적 새로운 팀입니까? Meituan은 왜 추천에 투자하기로 결정했나요?

저희 팀은 수년 전에 설립된 Meituan의 인프라 팀입니다.

검색, 추천 및 광고는 Meituan에서 가장 중요한 비즈니스 흐름이며 사용자 성장과 비즈니스 실현을 강화합니다.

Q: 팀에서 어떤 종류의 추천 시스템에 초점을 맞추고 있습니까?

추천 시스템의 경우, 저희 팀은 주로 모델 훈련 최적화에 초점을 맞추고 있습니다. 오프라인 및 온라인 특징을 유연하고 효율적으로 읽을 수 있는 효율적인 데이터 흐름 시스템을 설계했습니다. 저희는 수평 확장의 관점에서 시스템을 최적화하고 시스템을 거의 선형적으로 확장하고 있습니다. 또한 수직 확장의 관점에서 시스템을 최적화하고 하드웨어 리소스를 완전히 활용합니다. 그리고 시스템 및 알고리즘 관점에서 시스템의 훈련 성과와 정확도를 공동으로 최적화합니다. 계속해서 새로운 모델 및 새로운 하드웨어의 SOTA에 주목하고 가치 있는 기술을 Meituan의 비즈니스에 도입할 것입니다.

Q: Meituan의 사용자 수는 6억3000만 명이고 사용자당 상호작용 수도 많습니다. 팀은 작업을 어떻게 훈련하나요? 훈련 빈도는 어느 정도인가요?

우선 알고리즘 전략 시험을 위해 소수의 샘플을 사용합니다. 검증이 효과적이면 많은 수의 샘플을 사용하여 위 전략을 실험하고 더 나은 결과를 얻으려 합니다.

매일 다양한 실험을 하지만 온라인 모델의 경우 훈련 빈도는 하루에서 일주일입니다.

Q: 팀에서 추천 시스템을 어떻게 평가하나요? 미세 조정은 어떻습니까?

추천 시스템에서는 다양한 모델 구조, 알고리즘 전략 및 샘플 기능을 실험하여 일련의 모델을 생성하고 이러한 모델이 오프라인 및 온라인 평가를 통해 비즈니스 지표의 개선을 가져올 수 있는지 여부를 판단합니다. 일반적으로 더 큰 샘플과 더 복잡한 모델을 훈련하면 비즈니스 지표가 향상되지만 더 많은 리소스가 소요되고 실험 횟수가 줄어듭니다. 그 사이의 균형을 맞추고 성능을 대폭 높일 수 있는 경우 더 큰 샘플과 더 복잡한 모델을 훈련하는 것을 선호합니다.

Q: 추천 시스템은 어떻게 최적화하십니까?

처음에는 CPU 아키텍처를 기반으로 훈련 프레임워크를 최적화했지만, 모델이 점점 더 복잡해짐에 따라 훈련 프레임워크를 심층 최적화하기가 어려웠습니다.

지금은 A100 GPU를 기반으로 훈련 시스템에 NVIDIA HugeCTR을 통합하려고 작업 중입니다. 8개의 A100 GPU가 있는 단일 서버는 CPU 기반 훈련 시스템으로 일하는 수백 명의 근로자를 대체할 수 있습니다.

비용도 크게 절감됩니다. 이것은 예비 최적화의 결과이며, 앞으로 최적화할 여지는 많습니다.

Q: 작업을 지원할 적절한 기술, 패키지, 방법, 프레임워크는 어떻게 선택하시나요?

진보되고 개방적이며 생태적이어서 이를 바탕으로 내부의 요구를 더 잘 충족할 수 있는 기술이어야 합니다.

Q: 오픈 소스 기술과 팀의 업무에 대한 상호 운용성은 얼마나 중요합니까?

매우 중요합니다. 저희 팀은 현재 오픈 소스 기술을 주로 사용하여 시스템을 구축하고 있습니다. 그와 동시에 저희의 작업 결과를 오픈 소스 커뮤니티에 되돌려 줄 수 있어서 기쁩니다.

Q: 최근 Meituan은 2021년 2분기까지 12개월 동안 사용자당 거래가 평균 32배 증가했다고 보고했습니다. 팀은 모델 확장에 어떻게 대처하나요?

더 많은 샘플과 더 복잡한 모델을 사용하여 비즈니스 모델을 표현함으로써 비즈니스 효과를 크게 향상했습니다.

Q: 팀의 최근 성공 사례는 무엇입니까?

A100+HugeCTR을 기반으로 차세대 추천 시스템에 대한훈련육 프레임워크를 개발하고 예비 결과를 달성했습니다.

Q: 최근에 추천 시스템 워크플로우에 특정 방법, 기술, 플러그인, 라이브러리 또는 패키지를 통합했나요?

원격 분산 파일 시스템에서 로컬 CPU 메모리에 이르는 전체 데이터 흐름 파이프라인을 GPU 메모리로 만들어 계산 및 IO가 중첩될 수 있도록 했습니다. 임베딩 레이어와 덴스 레이어도 파이프라인이므로 임베딩 레이어는 더 이상 전체 시스템의 병목 구간이 아닙니다.

Q: 팀장이 이제 막 시작해서 현재 추천 시스템 구축, 배포, 최적화를 평가 중이라면 어떤 조언으로 추천 시스템 워크플로를 가속 및 간소화하는 데 도움을 주시겠어요?

현재 회사의 인프라와 비즈니스 상태를 충분히 이해하고 이를 바탕으로 시스템 및 프로세스를 설계하세요. 기술 스택과 프레임워크를 선택할 때 각 시스템의 성숙도, 커뮤니티 생태, 확장성 및 통합 친화성을 고려하세요.

FELIPE CONTRATRES

개인화 팀장

Magalu (Magazine Luiza)

Q: Magalue에서 어떤 직책을 맡고 계신가요? 무슨 일을 담당하고 계신지요?

저는 현재 Magalu 개인화 팀의 팀장을 맡고 있습니다. 제 역할은 Magalu의 비즈니스 목표와 일치하는 유용한 고품질 소프트웨어를 프로덕션 환경에 제공하는 것입니다.

Q: 팀은 Magalu에서 어떤 작업을 하나요? 팀의 책임은 무엇인가요?

저희 팀은 추천 및 검색 플랫폼을 구축하기 위해 노력하고 있습니다. 저희 팀에는 시스템 개선을 위해 노력하는 개발자와 데이터 사이언티스트가 있습니다. 저희 추천 팀은 플랫폼과 새로운 추천 모델을 위한 새로운 기능을 만들기 위해 노력하고 있습니다.

Q: Magalu의 전반적인 비즈니스와 팀의 추천 시스템 작업에는 어떤 관련이 있나요?

추천 시스템은 Magalu의 수익에서 상당한 액수를 차지하고, 사용자 경험에 크게 기여하며, Magalu의 여러 KPI에 많은 영향을 미칩니다.

저희는 많은 Magalu 채널 (데스크톱 웹 사이트, 모바일 웹 사이트, 앱, MagazineVocê 및 1400개 이상의 오프라인 매장 전부) 에 추천 시스템을 제공합니다. 또한 광고 팀 (Magalu Ads) 이 플랫폼에 추천 광고 알고리즘을 게시하도록 지원합니다.

Magalu 플랫폼은 월간 3200만 명의 활성 사용자에게 수백만 개의 제품과 서비스를 추천하고 있습니다.

Q: 팀은 비교적 새로운 팀입니까? Magalu는 왜 추천에 투자하기로 결정했나요?

Magalu는 2013년부터 추천 팀을 두고 있습니다. 당시에는 서드파티 기업에서 제공하는 추천 시스템을 사용했습니다. 하지만 추천 시스템이 마켓플레이스 플랫폼에서 핵심 포인트가 되면서 내부적으로 직접 추천 시스템을 만들기로 결정했습니다.

Q: 팀에서 어떤 종류의 추천 시스템에 초점을 맞추고 있습니까?

저희 팀은 제품 추천 시스템에 주력합니다. 다른 팀이 새로운 추천 모델을 만드는 데 협력할 수 있는 추천 플랫폼을 구축했습니다. 저희 팀은 플랫폼을 유지하고 비즈니스 팀에서 사용할 새로운 기능을 만드는 일을 맡습니다.

Q: 팀에서 사전 처리 (ETL) 및 특징 엔지니어링을 어떻게 처리하나요?

저희 데이터 플랫폼 팀은 모든 데이터 수집 및 사전 처리를 처리하는 데 사용하는 놀라운 내부 데이터 플랫폼을 만들었습니다. 또한 특징을 저장하고 모델을 훈련하는 데 사용하는 내부 머신 러닝 플랫폼도 있습니다.

Q: 팀에서 추천 시스템을 어떻게 평가하시는지요? 미세 조정은 어떻습니까?

새로운 모델에 대한 오프라인 및 온라인 평가를 사용합니다. 추천 플랫폼에서 누구나 두 가지 모델을 온라인에 배치하고 생산 효율을 비교할 수 있는 A/B 테스트 기능을 구축했습니다.

Q: 추천 시스템은 어떻게 최적화하십니까?

A/B 테스트 기능을 사용하여 온라인으로 테스트합니다. 모델의 새로운 버전을 만들어서 A/B 테스트로 온라인에 배포합니다.

Q: 작업을 지원할 적절한 기술, 패키지, 방법, 프레임워크는 어떻게 선택하십니까?

우선 추천 시스템 문헌을 살펴보면서 저희가 다루는 문제에 대한 과거의 연구가 있는지 찾아봅니다. 어떤 패키지 또는 프레임워크를 사용할지 결정하기 위해 커뮤니티가 얼마나 참여하고 프레임워크가 얼마나 잘 확립되었는지 항상 확인합니다.

Q: 모델 확장은 어떻게 해결하십니까?

Magalu는 복잡성을 대신 처리해 주는 훌륭한 머신 러닝 플랫폼을 보유하고 있습니다. 이 플랫폼을 사용하면 모델을 훈련하고 확장 가능한 방식으로 노출하여 인프라와 관련된 복잡성을 추상화할 수 있습니다.

Q: 최근의 성공 사례나 성공적으로 해결된 과제는 무엇인가요?

유연하고 확장 가능한 추천 플랫폼을 구축한다는 목표를 성공적으로 달성했습니다. 이 플랫폼을 통해 다른 팀이 새 전문화된 모델을 만들고 사용자 경험을 쉽게 맞춤 설정하여 기여할 수 있습니다.

Q: 최근에 워크플로우 통합을 고려하고 있는 특정 기술이나 방법에 대한 추천 시스템 논문이 있나요? 아니면 최근에 특정 방법을 통합했나요?

Gabiel de Souza P Moreira의 최근 논문, '전자 상거래 세션 기반 추천을 위한 다중 모달 특징 및 포스트 퓨전 맥락을 갖춘 트랜스포머 (Transformers with multi-modal features and post-fusion context for e-commerce session-based recommendation)' 논문에서 제안된 기술을 테스트하려고 고려 중입니다.

Q: 팀에서 세션 기반 추천 시스템을 평가하거나 사용하고 계신가요?

현재 세션 내 사용자 경험을 실시간으로 개인화하는 경험 세션 기반 추천 시스템을 보유하고 있으며, 고객 경험 향상을 위해 위에서 인용한 논문에 제안된 방법 (전자 상거래 세션 기반 추천을 위한 멀티 모달 피처 (Multi-modal features) 및 포스트 퓨전 컨텍스트 (Post-fusion context) 을 갖춘 트랜스포머) 를 평가하고 있습니다.

Q: 팀장이 이제 막 시작해서 현재 추천 시스템 구축, 배포, 최적화를 평가 중이라면, 추천 시스템 워크플로우를 가속 및 간소화하는 작업에 대해 어떤 조언을 주시겠습니까?

몇 가지 조언이 있습니다.

- 간단한 접근법부터 시작하세요. 많은 문제는 시스템 전반의 복잡도를 줄여 주는 간단한 기법으로 쉽게 해결됩니다.
- 고급 기법이 필요한 경우 워크플로우에 통합할 성숙한 패키지 또는 프레임워크를 선택하세요.
- 과도하게 오프라인에서 사전 최적화하는 대신 실제 고객과 함께 모델을 테스트하기 위해 가능한 한 빨리 모델을 온라인으로 전환하세요.

VINNY DEGENOVA

Wayfair 검색 및 추천 부서 데이터 사이언스 이사

Q: Wayfair는 어떤 기업인가요?

Wayfair는 가구 및 가정용품을 전문으로 하는 미국 전자 상거래 회사로 미국, 캐나다, 영국 및 독일에서 선도적인 입지를 점하고 있습니다. 현재 약 2200만 개의 제품 카탈로그를 보유하고 있으며 연매출은 148억 달러가 넘습니다.

Q: Wayfair에서 어떤 역할을 맡고 계십니까?

저는 Wayfair의 추천 부문에서 여러 역할을 수행했으며, 개인화된 추천을 전문으로 합니다. 지난 몇 년 동안 개인화된 제품 추천을 전문으로 하는 팀을 구축했습니다. 저는 현재 검색 및 추천 팀의 이사로서 콘텐츠 추천 데이터 사이언스 팀을 이끌고 있으며, 알맞은 콘텐츠 (판매 이벤트, 동영상, 영감을 주는 기사, 이메일 콘텐츠 등) 를 적절한 고객에게 추천하여 사이트 안팎의 여정을 촉진하는 데 주력하고 있습니다. 목표는 각 고객을 위해 독특하고 개인화된 쇼핑 경험을 만드는 것입니다.

Q: Wayfair에서 귀하가 이끄는 팀의 책임은 무엇인가요?

저희 팀은 Wayfair 웹 사이트, 앱, 마케팅 이메일, 푸시 알림 및 타사 웹 사이트에서 콘텐츠 추천을 제공하는 추천 알고리즘을 개발합니다. 엔지니어링 팀과 긴밀히 협력하여 이러한 시스템을 생산 및 확장하고, 하루에 수십억 개의 권장 사항을 제공하고, 분석 파트너와 긴밀히 협력하여 그 영향을 정량화하고, 다음으로 착수할 기회를 평가합니다.

Q: Wayfair의 전반적인 비즈니스와 팀의 추천 시스템 작업과는 어떤 관련이 있나요?

개인화 작업은 고객이 좋아하는 제품을 쉽게 찾도록 만들어 고객 경험에 지대한 영향을 미치며
Wayfair의 비즈니스에 측정 가능한 직접적 영향을 미칩니다. A/B 테스트를 활용하여 새로운 추천 모델의 성능을 명시적으로 측정할 뿐만 아니라 현장에서 배포하는 추천 시스템에 대한 업데이트도 영업 및 사용자 참여 측면에서 명시적으로 측정합니다.

Q: 팀은 비교적 새로운 팀입니까? Wayfair는 왜 추천에 투자하기로 결정했나요?

Wayfair의 추천 팀은 제가 입사하기 전부터 있었습니다! 저는 4년 반 전에 이 회사에 왔습니다. Wayfair는 전자 상거래 회사로서 고품질 추천에 투자하는 것이 중요하다는 사실을 오래 전부터 알고 있었습니다. 저희가 제품과 콘텐츠 추천에 사용하는 시스템은 시간이 갈수록 비교적 단순한 클릭 집계 및 협업 필터링에서 보다 탄탄한 딥 러닝, 그래프 기반, 강화 학습 솔루션으로 발전했습니다. 그러면서 고객에게 무엇을 보여줄지 생각할 때 최신 기술 스택을 확보하는 것이 Wayfair 비즈니스 모델의 핵심적인 부분이 됐습니다.

Q: 팀에서 어떤 종류의 추천 시스템에 주력하고 있나요?

저희 팀은 다양한 유형의 콘텐츠를 아우르는 추천 시스템을 개발하는 데 주력하고 있습니다. 여러 가지 접근법을 활용하여 다양한 목적에 맞는 추천을 생성하고 있습니다. 예를 들어, 현재 사이트에서 벌어지는 판매 이벤트에 대한 맞춤형 추천을 생성할 때는 강화 학습 기술 조합을 사용하여 어떤 판매 이벤트가 가장 인기인지 파악하고 고객의 이전 쇼핑 행동을 바탕으로 맞춤형 휴리스틱 (Heuristics) 을 개발하여 고객이 이메일에서 보게 될 판매 이벤트 목록을 조정합니다. 개인화된 제품 추천의 경우, 저희 팀은 Wayfair 웹 사이트에서 고객이 가장 많이 참여할 것으로 생각되는 제품 순위를 실시간 지정하는 딥러닝 기법을 구축하는 데 투자했습니다. 저희가 구축한 모델에 대한 상세 정보는 Wayfair 테크 블로그의 게시물을 여기서 참조하세요.

Q: Wayfair 고객은 한 번에 '하나의 항목, 하나의 방, 하나의 프로젝트를 쇼핑'하는 선호를 보였는데, 팀에서 이와 관련된 추천을 어떻게 제공하셨나요?

추천 시스템을 개발할 때 숙달해야 할 중요한 요소 가운데 하나는 추천 후보를 잘 이해하는 것입니다. 제품 추천의 경우 보통 Wayfair의 데이터 사이언스 관련 파트너 팀이 소유하며, 이 팀은 알고리즘에 따라 특정 스타일에 연관된 제품을 태그 지정하는 프로세스를 자동화합니다. 다른 유형의 추천의 경우, 저희는 피처 추출 전용 파이프라인 (Feature derivation pipeline) 을 구축하여 이미지, 텍스트, 기타 소스로부터 유의미한 정보를 추출하며, 이를 소비자의 브라우징 이력과 조합하여 Wayfair 내 고객의 모든 접점에서 가장 연관성 있는 콘텐츠를 추천할 수 있습니다. 이는 고객에게 추천할 수 있는 항목에 대해 깊이 파악할 수 있게 도와줍니다. 이와 동시에 고객이 구매하려는 항목을 파악하기 위해 Wayfair 웹 사이트와 인터렉션하는 방식을 분석하고, 추천을 활용하여 고객의 필요와 가장 연관성 있는 제품과 콘텐츠를 표시함으로써 구매 사이클에서의 마찰을 제거합니다.

Q: 팀에서는 사전 처리 (ETL) 및 피처 엔지니어링 (Feature engineering) 을 어떻게 처리하나요?

저희는 머신 러닝 플랫폼 팀과 긴밀하게 협력하여 특징의 재사용 가능성을 생각합니다. 특정 사용 사례에 대한 틈새 특징은 항상 있겠지만 추천의 경우 일반적으로 사용되는 다수의 특징이 여러 사용 사례에서 유사합니다. Wayfair는 실시간으로 추론을 실행하는 모델을 위해 특징을 생성하는 플랫폼 구축에 크게 투자했습니다. 이러한 특징은 주기적으로 업데이트되고 여러 프로젝트, 여러 팀에서 재사용 가능하며 오프라인 포맷 (보통 GBQ 테이블) 으로 액세스 가능하거나 온라인으로 액세스 가능하게 해 줍니다.

Q: 팀에서는 어떻게 훈련을 실시하나요? 훈련 빈도는 어떻습니까?

훈련 빈도는 활용 사례에 따라 다릅니다. 일반적으로 저희는 모델 조정 및 예약에 Airflow를 활용하며 Google Cloud를 사용하여 훈련 작업을 리소스로 배포합니다. 보다 예측적인 모델링을 위해 작업에 따라 짧으면 매일, 길면 매주 간격으로 훈련 작업을 실행합니다.

온라인 학습 시스템의 경우, 저희는 일괄 업데이트를 훨씬 자주 실시하며, 일간 업데이트를 통해 고객의 행동 변화에 최대한 빨리 대응하고자 합니다.

Q: 플랫폼의 활성 고객은 3100만 명이고 제품은 2200만 개가 있으며 매일 수십억 개의 추천이 이뤄지는데, 팀에서 추천 시스템을 어떻게 확장하나요?

확장성은 처음부터 모델 개발에 포함되어야 합니다. 이 크기의 카탈로그와 저희가 운영하는 규모의 고객 기반을 고려하면 모델 훈련, ETL 파이프라인 또는 모델 추론에 비효율의 여지가 허용되지 않습니다. 저희가 작업하는 모든 모델은 고객에게 관련 항목을 보여주는 측면 뿐만 아니라 확장성과 유지 관리 측면에서도 테스트 및 평가됩니다. 추천 사항을 생성하는 데 걸리는 시간 및 모델 교육에 걸리는 시간과 같은 정보를 명확하게 파악할 수 있기 때문에 모델이 실제 생산되기 전에 엔지니어링 팀과 긴밀하게 협력하여 비효율성에 대한 플래그를 지정하고 식별할 수 있습니다.

Q: 팀에서는 추천 시스템을 어떻게 평가하고 미세조정하나요?

온라인 A/B 테스트를 시작하기 전에 오프라인에서 제품 추천 시스템을 평가하고 미세 조정하는 인하우스 패키지 (In-house package) 를 개발했습니다. 과거 데이터에서 모델을 백테스트하고, 새 모델이 이 사이트에 현재 사용 중인 모델에 비해 얼마나 잘 작동할지 평가합니다. 일반적으로 Wayfair의 모든 검색 및 추천 이니셔티브에서 백테스트를 수행하며 가드레일 메트릭 (Guardrail metrics) (예: 리뷰가 없는 제품은 추천하지 않기) 에 크게 의존합니다. 또한 휴리스틱 (Heuristics) 을 활용하여 개발한 새 모델이 고객 경험을 개선하는지 확인합니다.

Q: 팀에서 최신 연구를 검토할 때 작업을 지원할 적절한 기법, 패키지, 방법, 라이브러리 또는 프레임워크를 어떻게 선택하나요?

Wayfair의 데이터 사이언스 팀에서 하는 작업을 '응용 연구'라고 말씀드리고 싶습니다. 다시 말해, 연구는 특정 비즈니스 문제에 응용됩니다. 해결하고자 하는 문제에 대한 문서화된 솔루션이 없을 때, 우리는 또한 최적의 접근 방식을 통해 생각할 수 있는 화이트보드로 이동할 수 있는 유연성을 가지고 있습니다. 영감을 찾으려 할 때나 선도적인 콘퍼런스에서 나온 최신 논문을 읽을 때 보통 Wayfair만의 영역에 응용하고 적용할 수 있는 기법을 모색합니다. 저희가 사용하는 프레임워크 또는 라이브러리는 저희가 추구하는 가치만큼 중요하지 않습니다. 구축하는 모든 솔루션을 카탈로그 및 고객 기반의 성장에 따라 관리 및 확장 가능하게 만드는 것이 중요합니다.

Q: 특정 방법, 패키지, 라이브러리 또는 프레임워크를 추천 시스템 워크플로우에 통합한 최근 사례는 무엇인가요?

최근 Google Cloud를 모든 데이터 웨어하우징, 모델 훈련, ETL 파이프라인, 모델 서비스에서 선호 플랫폼으로 활용하기 위해 온 프레미스 데이터센터로부터 대규모 마이그레이션을 마쳤습니다. Wayfair의 모든 엔지니어링 팀이 힘을 합친 결과지만, 데이터 사이언스 팀은 이미 모델 개발, 훈련, 배포에서 보다 효율적이고 확장성 뛰어난 클러스터로 인해 이득을 보고 있습니다.

Q: 팀장이 이제 막 시작해서 현재 추천 시스템 구축, 배포, 최적화를 평가 중이라면, 추천 시스템 워크플로우를 가속 및 간소화하는 작업에 어떤 조언을 주시겠습니까?

간단하게 시작하세요. 종종 비교적 간단한 모델로도 문제를 해결할 수 있는 여지가 상당히 많습니다. 아니면 보다 복잡하고 섬세한 솔루션으로 넘어가기 전에 간단한 구현으로 쉽게 얻을 수 있는 이득도 있습니다.

바퀴를 다시 발명할 필요가 없다는 말입니다. 새로운 유형의 추천 시스템 연구가 정기적으로 진행 및 발행되고 있지만, 추천 시스템 분야는 오래 전부터 있었습니다. 이는 모델을 가속화하기 위해 활용할 수 있는 오픈 소스 리소스가 많다는 것을 의미합니다. 모든 것을 처음부터 구축할 필요는 없습니다. 주변 커뮤니티를 최대한 활용하세요.

규모를 염두에 두고 개발하세요. 처음부터 새 모델의 확장성을 고려하지 않으면 프로덕션 환경으로 넘어갈 때 실패합니다. 앞으로 몇 주, 몇 달 또는 몇 년 동안 모델이 어떻게 확장될지 분명하게 파악한 상태로 개발 프로세스를 시작하는 것이 현명합니다!

KANNAN ACHAN

개인화 및 추천 팀장

Walmart Global Tech

Q: Walmart Global Tech에서 어떤 직책을 맡고 계십니까? 어떤 작업을 담당하고 계신지요?

Walmart Global Tech에서 개인화 및 추천 팀을 이끌고 있습니다.

Q: Walmart Global Tech에서 팀은 무슨 일을 하고 있는지요? 팀의 책임은 무엇인가요?

저희 팀은 고객 여정의 개인화를 책임집니다. 일반적으로 추천 시스템에서 이는 단일 추천 캐러셀입니다. 기본적으로 페이지 내의 단일 모듈입니다. 그러나 이 팀은 전체 페이지 개인화를 추구합니다. 그리고 모든 것은 한 번에 개인화 팀으로 들어옵니다.

Q: Walmart Global Tech의 전반적인 비즈니스와 팀의 추천 시스템 작업과는 어떤 관련이 있습니까?

개인화 시스템은 이메일부터 홈 페이지에서의 발견, 항목 페이지, 거래 후, 마케팅, 음성 상거래까지 고객 여정 도중에 50개 이상의 접점을 보유하고 있습니다. 저희는 전체 페이지를 개인화하지만 단순히 페이지만을 개인화하는 것으로는 충분하지 않습니다. 이를 어떤 면에서는 텐서 (Tensor) 로 생각하고 전체 세션도 개인화해야 합니다. 고객이 누구인지 아는 것이 가장 중요합니다. 고객이 사이트에 있을 때, 그들이 왜 있는지 파악하고 원하는 것을 발견하게 해야 합니다. “발견 단계 (Discovery funnel) 에서 추천이 활용되어 전환을 돕는 것입니다. 보다 중요한 것은 이 여정에서 저희가 고객을 다시 참여하게 만드는 모델을 작업한다는 사실입니다.

Q: 팀은 비교적 새로운 팀입니까? Walmart Global Tech는 왜 추천 시스템에 투자하기로 결정했나요?

제가 입사했을 때 처음 맡은 프로젝트 중 하나는 추천 시스템을 만드는 것이었습니다. 곧 저희는 고객 데이터에 대한 통합된 뷰를 제공해 달라는 요청을 받았습니다. 당시 Walmart 온라인, Walmart 매장, Sam's Club이 있었는데 이 모든 곳에서 많은 고객이 쇼핑을 하고 있었습니다. 목표는 통합된 고객 ID를 생성하고 고객의 쇼핑 방식을 파악하는 것이었습니다. 그게 6년, 7년 전의 일입니다. 그때부터 회사가 개인화에 투자하기로 결정했고 저희는 그러한 풍부한 고객 이해를 활용하여 홈 페이지에서 몇 가지 모듈을 개인화하기 시작했습니다.

Q: 팀에서 어떤 종류의 추천 시스템에 초점을 맞추고 있습니까?

추천 시스템 부문을 생각해 보면, 기존의 관점에서 추천 시스템은 사용자와 평점으로 구성된 행렬입니다. 즉 협업적 필터 (Collaborative filter) 죠. 발견 (Discovery) 을 일반적으로 살펴보면, '이 상품을 둘러본 고객은 다음과 같은 상품도 둘러보았습니다' 같은 전통적인 추천을 제공해야 합니다. 그러나 상품이 장바구니에 들어가고 나면 이를 보완하는 상품 또한 추천해야 하겠죠. 옴니채널 데이터가 복잡한 이유는 의도가 있기 때문입니다. 고객은 식료품을 구매할 수도 있고, TV 같은 일회성 구매를 할 수도 있습니다.

많은 고객이 식료품을 구매하러 Walmart에 옵니다. 고객의 의도가 식료품을 구매한다는 것을 예측한 순간 저희는 빠르게 전환하여 고객이 장바구니를 최대한 빨리 가득 채우도록 돕습니다. 장바구니 크기 외에도 빠른 거래 마무리를 최적화하는 다른 메트릭스도 있습니다. 저희 팀의 최신 발명은 장바구니 예측 기술입니다. 클릭 한 번으로 구매한 제품은 고객의 장바구니로 들어간 후 결제로 넘어갑니다.

Q: 작업을 지원할 적절한 기술, 패키지, 방법, 프레임워크는 어떻게 선택하십니까?

오프라인인 경우 데이터 사이언티스트가 훨씬 유연하게 모델을 선택하거나 개념 증명 구축을 선택할 수 있습니다.

그러나 40~50밀리초라는 제약 때문에 온라인에서 실행 가능한 것에는 상당한 제한이 있습니다. 대부분의 경우 이 제한이 기술 선택을 결정합니다. 때로는 아주 간편하게 단순한 로지스틱 회귀나 그레이디언트 부스팅 의사 결정 트리를 사용하기도 합니다. 이러한 방법도 잘 작동하고, 예측적으로 설명이 가능하며, 시간 지연을 아주 유의미한 방식으로 제어할 수 있기 때문입니다.

오프라인으로 실행되는 것과 온라인에서 실행되는 것 사이에서 절충점을 찾아야 하죠. 때로는 모델 크기가 커서 호스팅할 수 없습니다. 결국, 저희는 데이터 사이언티스트들에게 데이터에 맞는 모델을 만들 수 있는 권한을 부여했습니다. 저는 항상 "문제를 추측하지 말고 솔루션을 추측하라"고 말합니다.

Q: 모델 확장은 어떻게 해결하십니까?

분산 컴퓨팅은 매우 중요합니다. 저희도 이를 많이 활용합니다. GPU 같은 경우 온프레미스에도, 클라우드 플랫폼에도 서비스가 있습니다. 모든 것을 가능한 한 많이 컨테이너화하여 자동 확장이 원활하게 진행되게 합니다.

Q: 최근에 워크플로우 통합을 고려하고 있는 특정 기술이나 방법에 대한 추천 시스템 논문이 있나요? 아니면 최근에 특정 방법을 통합하셨나요?

잘 알려진 문제이기 때문에 탐색/활용에서 상당한 매력을 느끼고 있습니다. 상황별 밴디트 (Contextual bandits) 와 보상 모델 (Modeling of rewards) 을 사용하면 큰 이익을 얻을 수 있습니다. Exploit framework는 매출뿐 아니라 프리젠테이션 편향에도 도움이 됩니다. 이것이 비즈니스에 매우 유용할 수 있다는 점은 이미 수치로 증명된 바 있으며, 편향되지 않은 데이터 세트를 시도하기 때문에 저희 모델 또한 개선할 수 있습니다."

기존의 추천 시스템은 최상위 10개 후보만 표시했죠. 그래서 사용자는 최상위 10개 검색 결과와만 상호작용합니다. 최하위나 20위, 30위는 상위 제품이 품절되지 않는 한 결코 노출될 일이 없습니다. 그러나 모델이 탐색/활용을 수행할 경우, 특정 제품의 실적이 좋은 경우 온라인 학습을 계속하지만 분산도가 매우 높은 경우 이를 빠르게 표시하여 잘 작동하는지 확인합니다. 마찬가지로 특정 제품의 실적이 좋지 않지만 확실하지 않은 경우, 해당 제품을 더 많이 표시하여 제거할 수 있는지 확인합니다. 저희는 Thompson 샘플링의 변형을 활용하여 이를 아주 원칙에 입각한 방식으로 수행합니다.

유망한 연구가 많다고 여겨지는 영역 하나는 자기 지도 학습 (Self-supervised learning) 입니다. 고객이 사이트를 보면 저희는 누군가가 사이트를 봤다는 걸 압니다. 장바구니에 물건을 담는 데 10분, 결제까지 30분이 걸렸다면 그런 진행 상황도 알죠. 매장 데이터는 '고객이 이 항목을 구입했다'는 것인데, 그걸 구매한 이유와 항목의 순서는 알 수 없었습니다. 이전까지 팀은 옴니채널 데이터를 서로 연관지어서 두 데이터 소스를 연계해야 했습니다. 데이터 소스 하나는 장바구니에서, 다른 하나는 명확한 의도와 그 의도가 어떻게 행동으로 전환되는지에 대해서 말입니다. 그런 다음 변경 시의 스토리라인 데이터와 함께 전자 상거래 데이터에 대한 풍부한 이해를 가질 수 있습니다. 이 두 소스를 결합하고 정규화하는 것은 큰 문제입니다.

Q: 현재 어떤 역할을 맡고 계신가요? 어떤 작업을 담당하고 계신지요?

EVEN OLDRIDGE

Merlin 엔지니어링 선임

NVIDIA

저는 Merlin 엔지니어링을 이끌고 있습니다. Merlin은 추천 시스템을 위한 NVIDIA의 솔루션입니다. 딥 러닝 기반 추천 시스템에 중점을 두고 시작됐습니다. 제가 큰 관심을 갖는 분야이기도 합니다. 제가 참여한 이유 중 하나는 딥 러닝 기반 추천 시스템을 GPU에서 잘 작동하게 만드는 방법을 알아보는 것입니다. 당시 커뮤니티에는 추천 시스템이 GPU 하드웨어에 맞지 않는다는 인식이 있었습니다. 어느 정도는 사실이었습니다. 3~4년 전 당시 하드웨어에는 메모리가 별로 없었습니다. 11기가 정도면 최상급이었고 추천 시스템은 상당히 메모리가 제한됐습니다. 사용자와 항목을 나타내는 임베딩 (Embedding) 이 추천 시스템 모델 대부분을 구성합니다.

딥 러닝 프레임워크에 대해 생각해 보면, 이는 컴퓨터 비전의 시대에 구축되었습니다. 따라서 이러한 워크플로우 유형의 가속화가 우선시됐으며, 지금도 많은 경우에 그렇습니다. 그러나 프레임워크는 추천 시스템이 일반적인 사용 사례라는 생각으로 바뀌기 시작했습니다. 주요 프레임워크인 TensorFlow와 PyTorch를 각각 개발한 두 회사의 비즈니스 모델은 추천에 기반을 두고 있으므로, 이 회사들은 그 중요성을 알고 있습니다. 내부적으로 많이 하고 있지만, 그 도구는 본래 추천을 목적으로 개발된 것이 아니었습니다.

저희 팀은 이런 도구를 추천 분야에 맞게 개량하고 있습니다. 저희가 구축한 가속 데이터 로더 등 그중 일부는 프레임워크 내에 있지만, 점점 추천 시스템을 이루는 구성 요소 전체를 포함하는 방향으로 확장하고 있습니다.

Q: 팀이 딥 러닝에서 '더 폭넓은 플레이'로 어떻게 전환하고 있나요?

저는 개별 기여자로 시작하여 GPU에서 Python 데이터 사이언스 에코시스템을 수행하기 위한 프레임워크인 RAPIDS™와 딥 러닝 프레임워크 사이의 격차를 해소하기 위해 최선을 다했습니다. 그래서 NVIDIA에서 처음 6개월 동안 그 가교를 만들었습니다. 이 과정에서 PyTorch에 대한 데이터 로더 (Data loader) 를 만들 수 있었는데, PyTorch에서 딥 러닝 훈련을 약 15배까지 가속화했습니다. 그 개별 작업 하나가 성장으로 이어졌습니다. 엔지니어 한 명을 데려와서 함께 일했고, 제가 프로젝트를 주도했습니다. 그 뒤로 2명이 4명이 되고, 4명이 8명이 되고, 현재는 11명으로 늘어났으며 새 구성원을 모집하고 있습니다. NVTabular 쪽은 ETL과 훈련 프레임 워크 가속화에 중점을 두었습니다. 그리고 약 1년 후, Jensen이 중국에서 새로운 프레임워크를 구축하는 추천 시스템 프레임워크 팀인 HugeCTR과 파트너십을 맺었습니다. 그게 Merlin의 탄생이었습니다.

"HugeCTR 팀의 작업은 NVIDIA가 보다 수준 높은 추천 시스템에 주력하고 있음을 보여주는 대표적 사례입니다. 현재 이 팀이 작업 중인 프로젝트는 100테라바이트 모델입니다. 현재 그 어떤 NVIDIA 고객보다도 큰 규모지만, 고객들도 머지않아 이 수준에 도달할 것입니다. 그리고 그 규모에 도달하면 아주 스마트하게 생각해야 합니다. 이 팀은 어떤 팀이 오픈 소스에서 작업하고 발행하고 있는지 보고 이를 제품에 통합하고 있으며, 다른 팀에 지침도 제공합니다. 일부 팀은 이 기술을 자신들의 스택에 통합하여 채택합니다. 이 수준에서는 많은 작업이 큰 규모에서 이뤄지며 여기에 주력하는데, 그럴 만합니다. 아주 많은 비즈니스가 있으니까요. 추천 시스템으로 작업하는 기업의 규모는 굉장히 다양합니다. 그리고 큰 규모의 기업에서 요구되는 컴퓨팅 수준은 작은 기업과 비교가 되지 않습니다.

하지만 이제 이 여정을 완전히 바꾸기 시작했습니다. 제가 처음 시작할 때는 딥 러닝 기반 추천 시스템이 전부였지만 이제는 팀이 어떻게 좋은 추천 시스템을 구축할 수 있을지 고민하고 있습니다. 이를 어떻게 쉽게 만들 수 있을까요? 최종 솔루션만이 아니라 각 단계에서 솔루션을 어떻게 제공할지 생각합니다. 팀원 중 누군가는 이를 오마카세, 셰프의 테이스팅 메뉴라고도 부릅니다. '저희가 추천하는 구성 요소는 이런 것들입니다. 이것이 셰프 추천 메뉴입니다. 물론 다른 메뉴를 선택할 수도 있습니다. 하지만 저희는 전채 요리와 메인 요리, 샐러드, 수프, 디저트가 필요할 거라고 생각합니다. 그리고 거기에 어울리는 와인도 준비되어 있습니다.' 이런 식인 거죠. 저희 도구를 사용하여 추천 시스템을 구축하는 방법을 보여주는 포괄적 솔루션은 추천 시스템을 구축하려는 팀에 탄탄한 토대를 제공할 것입니다. 무엇보다 중요한 것은 이를 프로덕션 환경에 구현하여 더욱 발전시킬 수 있다는 사실입니다.

Q: 팀은 어디에 초점을 두고 있습니까

팀은 세 가지 요소에 중점을 두고 있습니다.

첫 번째는 추천 시스템을 더 쉽게 구축할 수 있도록 하는 것입니다. 저희는 특징 엔지니어링을 위한 도구를 만들고 있습니다. 모델링을 쉽게 만들고, 추천 모델을 쉽게 프로덕션 환경에 맞게 준비하는 데 필요한 도구를 제공하려는 것입니다.

두 번째는 이 모델을 어떻게 프로덕션 환경에 도입하느냐는 것입니다. 이것이 많은 회사에게 아주 큰 과제입니다. 이는 쉽지 않은 과제입니다. 딥 러닝 측면에서는 특히 그렇습니다. 기초적인 협업 필터링조차 뭔가를 프로덕션 환경에 구현하고 적절하게 모니터링하여 궤도에서 벗어나지 않게 해야 하고, 추적해야 하는 모든 것들을 제대로 추적하고 있는지 확인해야 합니다. 대단히 복잡합니다. 두 번째로 팀이 작업하는 것은 실운영에 쉽게 배치할 수 있도록 하는 것입니다.

그리고 세 번째는 GPU에서 성능을 발휘하는지 확인하는 것입니다. NVIDIA의 많은 팀들이 이를 가장 우선시하고 있습니다. 저희가 구축하는 것은 모두 GPU에서 수행됩니다. 그게 저희뿐 아니라 모두에게 가장 중요한 과제입니다. 제 생각에는 도구를 만드는 것이 중요합니다. 많은 경우 프로덕션 환경에 배포하는 것이 파이프라인에서 가장 어려운 부분이자 가장 큰 기회일 것입니다. 저는 이것이 정말로 우리의 초점이 옮겨가는 부분이라고 생각합니다. 굉장히 어렵기 때문에 쉽게 만들수록 고객들이 더 많은 성공을 거둘 것입니다.

딥 러닝과 덜 복잡한 ML 모델의 관점에서 볼 때, 그것은 많은 회사들이 이루고자 하는 여정입니다. 저도 그렇게 하려고 여러 번 시도했습니다. 다른 사람들과 함께 노력했죠. 이야기도 많이 나눴습니다. 늘 문제가 되는 것은 데이터 사이언티스트가 데이터 세트를 얻고, 모델을 구축하고, 그게 오프라인에서 좋아 보이면 흥분해서 이를 시도해 보려고 한다는 것입니다. 그 후 프로덕션에 모델을 구축할 때 넘어야 할 산이 있습니다. 페이스북처럼 세계적인 대기업에서는 이를 대규모 엔지니어링 팀과 함께 해결합니다.

Q: 모델링의 딥 러닝 측면 외에도 다른 특정 집중 분야가 있습니까? 예를 들어 세션 기반 추천 시스템에 대해 이야기하는 사람들이 많습니다.

세션 기반 추천 시스템은 아주 흥미로운 분야입니다. 추천 시스템 분야에서 약간 새로운 기술입니다. 특히 아시아 태평양 지역에서 많은 회사들이 사용하고 있는데, 이 회사들은 지난 수년 동안 ACM RecSys에서 상당한 이득과 성과를 보여주고 있습니다. 그들은 세션 기반의 훌륭한 모델들을 만들어냅니다. 그리고 지금 프로덕션 환경에서 실행하고 있습니다. RecSys 논문에서 상당한 성능 향상을 입증했습니다.

올해 저희 팀은 ACM RecSys에 이 분야에서 도움이 될 Transformer4Rec 라이브러리에 대한 논문을 제출하려고 준비 중입니다. 이 라이브러리의 목표 또는 개발 동기는 많은 추천 시스템 아이디어, 특히 세션 기반 추천 시스템이 NLP 영역에서 시작된다는 것이었습니다. 그러나 NLP 분야와 추천 시스템 분야 사이에는 간극이 벌어지고 있습니다. 추천 시스템보다 NLP에서 두 자릿수는 더 많은 연구들이 진행되고 있습니다. 현실적인 프로덕션 환경에서 컴퓨팅 양은 그 반대에 가깝다는 사실에도 불구하고 말입니다. 이는 대체로 데이터 세트가 부족하기 때문입니다.

Q: 팀을 이끌고, 고객과 협업하고, 문제를 심층 분석한 경험으로 미루어볼 때, 방금 시작한 사람에게 어떤 조언을 해 주시겠어요? 백지 상태에서 시작하여 자체적인 추천 시스템을 구축하려는 사람들이 처음 발을 들이려면 어떻게 해야 할까요?

간단하게 시작해서 반복 작업하는 것이 핵심이라고 생각합니다. 데이터 사이언티스트들은 가장 훌륭한 최신 딥 러닝 모델을 구축하고자 합니다. 보다 단순하고 간단한 것부터 시작해서 가능성이 있는지 확인합니다. 제 경험상 좋은 데이터는 항상 좋은 모델보다 낫습니다. 데이터 사이언티스트는 종종 험난한 경험 끝에 그런 교훈을 얻습니다. 저는 이것이 매우 보편적인 일이라고 생각합니다. 데이터를 잘 정리하고 관리한 다음 피처 디스커버리 (Feature discovery) 을 하면 최신 모델을 사용하자는 관점으로 시작하는 것보다 더 나은 결과로 이어질 가능성이 높습니다.

프로세스를 파악하는 것이 중요합니다. 그러면 '여기 모델이 있고, 반복 작업을 이렇게 해서 출력으로 이런 유형의 모델을 얻을 것'이라는 식의 관점에서 벗어날 수 있습니다. 그 모델을 프로덕션 환경에 구현할 때는 프로덕션 쪽에서 일하는 팀과 먼저 대화를 해서 무엇이 사용되고 어떻게 배포될 수 있는지, 제한 사항은 무엇인지 이해해야 합니다.

저희는 Merlin을 통해 간단한 모델로 시작하여 시간이 지남에 따라 더 복잡하게 구축할 수 있도록 프레임워크를 제공하려 하고 있습니다. 이 분야에는 빈틈이 있습니다. 현재 기존 머신 러닝에서 딥 러닝으로 전환하기가 굉장히 어렵습니다. 많은 회사들이 이 단계에서 고전하고 있으며, 딥 러닝 기반 작업을 수행하려면 완전히 다른 파이프라인을 구축해야 하는데 그게 아주 복잡합니다. 그 생애 주기와 MLOps, 반복 작업을 생각하면서 모델을 1~3개 구축할 때까지 어떻게 빠르게 반복 작업을 할지 알아내야 합니다. 시스템을 구축하고 모니터링을 하고 모델을 잘 이해하고 추적하고 데이터 신호를 이해하고 데이터가 깨끗하다는 것을 알 수 있다면 말입니다. 그런 점에서 모델링은 최적화에서 마지막 단계 중 하나지만 사람들은 종종 여기서부터 시작합니다.

더 간결하게 답변하자면, 저희는 Merlin이 좋은 시작점이 되기를 바라며 이를 위해 열심히 노력하고 있습니다. 추천 시스템을 개발하고 배포하는 것은 매우 복잡한 작업입니다. 물론 ML 전반이 그렇지만 추천 시스템은 더욱 그렇습니다. 하지만 이제는 변하고 있습니다. 저희는 사용과 구축이 쉽고 성능이 뛰어난 프레임워크인 Merlin에 총력을 기울이고 있습니다.

저자 정보

Paco Nathan

Derwen, Inc. 총괄 파트너

데이터 사이언스, 클라우드 컴퓨팅, 자연어, 그래프 기술의 핵심 전문성을 보유한 '선수/코치'로 알려져 있습니다. Bell Labs부터 초기 단계 스타트업까지 다양한 곳에서 40년 동안 IT 산업 경력을 쌓았습니다. Amplify Partners, Recognai, KUNGFU.AI 고문 PyTextRank, kglab 선임 커미터 이전: 이사, Community Evangelism @ Databricks 및 Apache Spark. 2015년 Innovation Enterprise에 의해 빅 데이터 및 분석 분야의 30대 인물로 선정.

자세히 알아보기

NVIDIA Merlin에 대해 보다 자세히 알아보려면 다음 웹 사이트를 방문하세요. developer.nvidia.com/nvidia-merlin

© 2021 NVIDIA Corporation. All rights reserved. NVIDIA, NVIDIA 로고, Merlin 및 [기타 NVIDIA 제품 및 기술]은 미국 및 다른 나라에서 NVIDIA Corporation의 상표 및/또는 등록 상표입니다. 다른 회사 및 제품 이름은 연관된 각 회사의 상표일 수 있습니다. 기타 모든 상표는 해당 소유자의 자산입니다.
SEP21