Lecture 15

Ksenia
Kasianova

Intro

Content
analytics

Web
Content
Monetiza-
tion

Search and
promotion
of content
in the
product

Technical
structure
of the
movie
hosting
website

Key
Business
Metrics

Dendrograms

End-to-
End
Analytics

Clustering
methods:

Lecture 15:
A/B testing in Content Analytics

Lecturer: Ksenia Kasianova
xeniakasianova@gmail.com

March 12, 2024

Plan

1) Content Analytics

- Web-content monetization

- Content search and promotion in the product.

2) Basic Analytical Concepts

- Technical part

- Basic metrics

- Hypothesis

- Dashboards

3) AB testing

- Problems in data

- Cause-and-effect relationships

- ML

4) Continuous A/B testing

Content analytics

Content analytics is the process of extracting meaningful insights and patterns from various types of content, such as text, images, audio, and video.

Content Analytics Examples:

Industry	**Subject**	**Task**	**Goal**
Marketing	Customer reviews, social media posts, website content	Analyze customer sentiment, preferences, and trends	Help marketers tailor their campaigns, create personalized content, and identify new opportunities
Media and Publishing	News articles, blog posts, social media discussions	Identify trending topics, key influencers, and public sentiment	Curate relevant content, optimize editorial strategies, and engage with audience
Customer Service	Customer support tickets, call transcripts, chat logs	Identify common issues, customer satisfaction levels, and agent performance	Improve response times, identify training needs, and enhance overall customer experience
E-commerce	Product descriptions, customer reviews, sales data	Understand customer preferences, purchase patterns, and product performance	Optimize product offerings, pricing strategies, and marketing campaigns
Healthcare	Medical records, research papers, patient feedback	Identify patterns, trends, and potential treatment options	Improve diagnosis accuracy, personalize treatment plans, and conduct research studies

Movie hosting industry

Content analytics in the movie hosting industry is the process of analyzing and
measuring the effectiveness of the content provided on a movie hosting platform.

Task: to understand what content is most popular with users, how they interact with
it, and what factors influence their behavior.

Goal: Content analytics helps determine the success of content, identify trends, and
predict future user needs.

Examples of content analytics in the movie hosting industry include:

1. **View Analysis:** Measuring the number of views of each movie or series on the
platform.

Goal: to determine the popularity of content and identify trends in user preferences.

Example: if a film receives a large number of views, this may be a sign of its success.

2. **Ratings analysis:** Studying the ratings left by users for each movie or TV series.

Goal: to understand how users evaluate content and what factors influence their preferences.

Example: if a film receives a high rating, this may be a sign of its quality.

3. **Watch Time Analysis:** Measuring the time users spend watching each movie or TV series.

Goal: to determine how interesting the content is to the user and how long they are willing to view it.

Example: if users spend a lot of time watching a particular series, this could be a sign of its compelling storyline.

4. **User behavior analysis:** Studying user actions on the platform, such as adding movies to a watch list, leaving comments or recommendations to other users.

Goal: to understand how users interact with content and what factors influence their decisions.

Example: if many users add a certain movie to their watch list, this could be a sign of its popularity.

5. **Conversion Analysis:** Measuring the number of users who, after watching a particular movie or TV series, subscribe to the platform or take other actions, such as purchasing a subscription or recommending to friends.

Goal: to determine the effectiveness of content in attracting new users and retaining existing ones.

Example: if a movie attracts a large number of new users, this may be a sign of its appeal.

Content Analytics Step-by-Step

1) Formulation of the problem

Example: create a service that suggests nearby restaurants, recommends certain products, or ranks movies or music that are interesting to users.

a) Identifying possible applications

- How many users will the resulting solution have?

- Who are they?

- Why do they need this functionality

- How will they find out about it?

b) Formulation of business metrics AND choice of the ones suitable for optimization in the process of training or selecting models.

The choice of metric depends on the specific task and objectives of the content analytics application:

- DCG-like metrics (Discounted Cumulative Gain) are commonly used in information retrieval tasks, such as ranking search results or recommendations.

Idea: more relevant items should be ranked higher in the list.

- Based on relative utility of relevant documents

- most useful at top ranks

$$DCG_k = \sum_{r=1}^{k} \frac{rel_r}{\log(r + 1)}$$

$rel_r = 11 \ldots 100 \ldots$ – relevance, r – ranks/order

Normalized version (NDCG):

- divide by DCG of best possible ranking:

- gives more weight to items at the top.

- AUC-like metrics (Area Under the Curve) are commonly used in binary classification tasks, such as sentiment analysis or spam detection.

AUC-like metrics measure the ability of a model to distinguish between positive and negative instances.

E.g.: Area Under the ROC Curve (AUC-ROC) – measures the overall performance of a classification model across different thresholds.

AUC-ROC is often used to evaluate the performance of sentiment analysis models or spam filters.

c) Decision of forming a training sample.

- What data is needed to form it, and how to obtain it?

- What is the event for our learning?

- Is sampling required? If yes, then how to do it?

d) Consideration of the cold start problem.

E.g.: when you already have the search engine, you can use user actions to receive signals about which documents are relevant to queries.

But what to do if the service has not yet been created, and the task being solved is critical for its functioning?

Challenge: making accurate predictions or recommendations when there is limited or no historical data available for these new items or users.

Methodologies:

1. Content-based Filtering: recommends based on the characteristics or attributes of the items themselves

How? Analyzes the features or metadata of the items and matches them with the user's preferences or past interactions.

Why? does not rely on historical data but rather on item attributes.

2. Collaborative Filtering: recommends based on the preferences or behavior of similar users.

How? looks for patterns or similarities in user-item interactions and uses this information to make predictions for new users.

Why? leverages the preferences of existing users to make recommendations for new users.

3. Knowledge-based Approaches: recommends based on the domain knowledge or expert rules

How? predefined rules or models that capture the relationships between items and users.

Why? no need for historical data.

4. Active Learning: actively seeking feedback or information from users to improve recommendations.

How? collect feedback from new users

Why? way to solicit their preferences explicitly.

5. Contextual Information: based on info on location, time, social connections, etc.

How? considering the context in which the recommendations are made

Why? more relevant and personalized predictions for new items or users.

2) Machine learning methods

a) Choose a model that will build a solution

- Which loss functional is optimized during the process of building a model and why it is a good choice for optimizing the chosen metrics

- Consider the optimization method used.

b) Feature space

Divide features into several classes according to the types of data used

– depends only on the user:

1. User demographics: Age, gender, location, etc.

2. User preferences: Previous movie ratings, genre preferences, and watch history $=>$

3. User activity: The frequency of movie watching, time spent on the platform, and engagement metrics $=>$ level of interest and commitment.

– depends on the "user - object" pair

1. Movie genre: a user who enjoys action movies may be more likely to watch an action-packed film.

2. Movie ratings: the average rating given by users can influence the recommendation for other users who have similar tastes.

3. Movie popularity: the number of views, likes, or positive reviews a movie has received can indicate its popularity among users

4. Movie release date: recommending new releases or trending films to users who prefer staying up-to-date

5. Movie actors or directors: involvement of specific actors or directors can impact its recommendation to users who are fans of their work.

3) Quality control

E.g. If a new solution was preceded by a previous one, is it of better quality?

a) Formulate experiments to test relevant hypotheses.

b) Choose an experimental model and a method for testing the statistical significance of changes.

Q:

- What indicators need to be tracked?

- What and how should you monitor to ensure the correctness of the experiment?

A:

- a regular A/B experiment on users of the service

- an expert assessment of the results of the work.

Web content monetization is the process of generating income from the creation and distribution of content on the Internet.

1. Advertising: contextual advertising, banners, video advertising and other advertising formats.

e.g.: the website owner gets paid for every ad click or view.

2. Affiliate Marketing: promoting the products or services of other companies on your website.

e.g. placing special affiliate links and receiving a commission for every sale or action made through those links

3. Paid Content: access additional content or benefits, like exclusive articles, video tutorials, content subscriptions, etc.

4. Sponsorship: receiving money from sponsoring companies for individual articles, videos or the entire website posting their advertising material in return

5. Selling Products or Services

e.g.

- paid consultations or training courses, one-on-one consulting, webinars, online

Lecture 15

Ksenia
Kasianova

Intro

Content
analytics

Web
Content
Monetiza-
tion

Search and
promotion
of content
in the
product

Technical
structure
of the
movie
hosting
website

Key
Business
Metrics

Dendrograms

End-to-
End
Analytics

Clustering
methods:

Web Content Monetization

Sources of profit for movie hosting:

1. Subscriptions:

– different levels of subscriptions, such as monthly, annual, or one-time payments for access to certain content.

– different subscription plans, including basic, standard, and premium, with different level of content access.

2. Advertising:

– promotional videos before the start of a movie

– display banners

3. Affiliate Programs, Sponsorships and Partnerships:

– partnership agreements with movie studios or content providers $=>$ commissions on every sale or rental of content made through their platform.

– partnerships with brands or companie: posting brand promotional materials on the platform, sponsoring certain films or TV series, or creating collaborative content or events.

4. Content Licensing:

– selling the rights to show films and TV series to other companies => expanding the audience and generate additional income

5. Partnerships with movie theaters:

– exclusive ticket deals or discounts on movie admissions

– watch new releases in theaters through their platform

6. Partnerships with device manufacturers:

– pre-installing their platform on TVs, smartphones or other devices

– earn income from the sale of devices or receive commissions from each view of content on those devices

Main goals:

1. Customer attraction AND retention

2. Improved user engagement

- Personalized recommendations

- What types of movies users enjoy, how they interact with the platform, and what keeps them coming back?

3. Content curation and catalog management

- Identify high-quality or popular movies that should be highlighted

- Identify any gaps in the catalog that need to be filled.

4. Content acquisition and licensing decisions:

- Which movies to acquire or license to meet user demand and maximize revenue?

5. Content performance analysis

- Decisions regarding promotion, marketing, or content optimization.

6. Fraud detection and content moderation:

- Fake reviews or illegal content uploads.

The stages of searching and promoting content:

1. Market Research and Analysis:

- understand the needs and preferences of the target audience based on study of competitors, trends in the entertainment industry

2. Content Planning:

- determine what content will be offered, in what genres, and how the library of films and TV series will be organized.

3. Search and acquire content

- making deals with film companies, studios, producers or distributors to obtain the rights to show content

4. Organize and index content into a user-friendly structure

- create sections and categories for different genres of films and TV series, ability to search by actors, directors and other parameters.

5. Recommendations and Personalization:

- based on user browsing history, ratings and reviews

6. Marketing and promotion:

- use of channels such as social media, advertising campaigns, partnerships with other brands or celebrities

7. Analysis and content optimization:

- adjust content strategy, recommendation algorithms, or product functionality based on this data

The technical structure of a movie hosting website may include the following components and functionality:

1) Frontend: content display, navigation and interactivity.

- Home Page: previews of new movies and TV series, popular recommendations, promotions and other content information.

- Movies and TV series pages: detailed information about films and TV series, including descriptions, ratings, cast, trailers and reviews.

- Search

- Categories and Genres

- Registration and Authentication: accessing additional features such as saving favorites, rating movies and TV series, leaving reviews, etc.

2) Backend: processes user requests, manages the database and provides the functionality of the site.

- **Server-based architecture:**

1. Centralized Architecture

Single server: Hosts all movie files and serves them to clients.

Advantages: Simple to implement + centralized control over content.

Disadvantages: Performance bottlenecks during peak traffic + Single point of failure.

2. Distributed Architecture

Multiple servers: Store and serve movie files across a network.

Advantages: Improved performance by distributing load + Increased availability and redundancy.

Disadvantages: More complex to implement and manage + Requires coordination between servers.

3. Cloud-Based Architecture

Cloud providers (e.g., AWS, Azure): Host and manage movie files on their infrastructure.

Advantages: Scalability and elasticity on demand + Reduced infrastructure costs and management.

Disadvantages: Potential for higher latency and bandwidth costs + Limited control over infrastructure.

4. Content Delivery Network (CDN)

Network of geographically distributed servers: Cache and deliver movie files closer to clients.

Advantages: Improved performance by reducing latency + Increased availability and reliability.

Disadvantages: Additional costs + Potential for content leakage.

- **Database management**

1. The vast amount of data associated with movie hosting, including:

Movie metadata (e.g., titles, descriptions, genres, release dates)

User information (e.g., profiles, preferences, watch history)

Streaming data (e.g., playback statistics, bandwidth usage)

Financial transactions (e.g., purchases, subscriptions)

2. Database Design:

Relational database: Most commonly used for movie hosting, allowing for structured data storage and efficient querying.

NoSQL database: Can be used for specific use cases, such as handling large volumes of unstructured data (e.g., user reviews).

3. Data Model:

- Entity-Relationship (ER) diagram: Defines the relationships between different entities in the database (e.g., movies, users, genres).

- Normalization: Ensures data integrity and reduces redundancy by organizing data into tables based on their relationships.

4. Indexing and Optimization:

- Indexing: Improves query performance by creating additional data structures that allow for faster data retrieval.

- Caching: Stores frequently accessed data in memory for faster retrieval.

- Partitioning: Divides large tables into smaller, more manageable chunks to improve performance and scalability.

5. Data Management Tasks:

- Data ingestion: Importing movie metadata and other data into the database.

- Data cleansing: Removing duplicate or invalid data to ensure data quality.

- Data updates: Handling changes to movie information, user profiles, and other data.

- Data analytics: Analyzing data to gain insights into user behavior, content performance, and other metrics.

6. Database Administration:

- Backup and recovery: Ensuring data protection and availability in case of system failures.

- Performance monitoring: Tracking database performance and identifying bottlenecks.

- Security: Implementing access controls and encryption to protect sensitive data.

7. Additional Considerations:

Scalability: Choosing a database that can handle the growing volume of data and user traffic.

High availability: Ensuring that the database is always available to users, even during maintenance or outages.

Integration: Integrating the database with other systems, such as content delivery networks (CDNs) and user interfaces.

- API (Application Programming Interface): provide an API that allows other applications or services to interact with data on the site.

- User management: handling user requests related to registration, login, password recovery and user account management.

- **Infrastructure:**

Operating a movie hosting website requires an infrastructure that ensures its availability, scalability and security.

This may include:

- Hosting: The site can be deployed on servers that ensure its operation and accessibility to users.

- Load Balancing: If a site has a large number of users, load balancing can be used to distribute requests evenly across multiple servers and ensure high performance.

- Caching: To improve site performance, you can use caching to save frequently requested data and reduce the load on servers.

- Security: The website must be protected from hacks and intruders. This may include DDoS protection, data encryption, user authentication and authorization, and other security measures.

Formulating the hypothesis

- User behaviour

Hypothesis 1: Users who watch movies on the platform on a mobile device are more likely to churn.

Hypothesis 2: Users who watch movies on the platform in the evening are more likely to engage with the platform.

Hypothesis 3: Users who watch movies on the platform with subtitles are more likely to be international users.

Hypothesis 4: Users who watch movies on the platform in high definition are more likely to be satisfied with the platform.

Hypothesis 5: Users who watch movies on the platform with a premium subscription are more likely to be loyal to the platform.

- Main drivers of user behavior:

Hypothesis 6: The more personalized the movie recommendations are, the more likely users are to watch movies on the platform.

Hypothesis 7: The more exclusive the content is, the more likely users are to subscribe to the platform.

Hypothesis 8: The more affordable the platform is, the more likely users are to subscribe to it.

Hypothesis 9: The more convenient the platform is, the more likely users are to use it.

KEY IDEA: formalize product requirements in the form of indicators and their changes

Measuring growth and engagement:

1) Monthly Active Users (MAU): The number of unique users who have engaged with a product or service within the past 30 days.

2) Daily Active Users (DAU): The number of unique users who have engaged with a product or service on a given day.

- A high DAU relative to MAU indicates a high level of user engagement.

- MAU and DAU: Can be calculated from user login data or activity logs.

- MAU and DAU provide insights into the size and activity level of a user base.

Measuring growth and engagement:

Metric	Formula
Monthly Active Users (MAU)	Number of unique users who have engaged with a product or service within the past 30 days
Daily Active Users (DAU)	Number of unique users who have engaged with a product or service on a given day
Weekly Active Users (WAU)	Number of unique users who have engaged with a product or service within the past 7 days
Average Daily Active Users (ADAU)	Total DAU over a period of time / Number of days in the period
Average Monthly Active Users (AMAU)	Total MAU over a period of time / Number of months in the period
Engagement Rate	(Total number of engagements / Total number of users) * 100%
Time Spent per Session	Total time spent by users on a product or service / Number of sessions
Pages per Session	Total number of pages viewed by users / Number of sessions
Bounce Rate	Percentage of users who leave a website after viewing only one page

Estimating revenue potential:

3) Lifetime Value (LTV): The total revenue that a customer is expected to generate over their lifetime.

- LTV: Can be estimated using predictive analytics based on factors such as customer demographics, usage patterns, and purchase history.

- helps businesses forecast future revenue and make informed decisions about customer acquisition and retention strategies.

Estimating revenue potential:

Metric	Formula	Description
Customer Life-time Value (CLTV)	(Average Purchase Value) * (Average Purchase Frequency) * (Customer Lifespan)	The total amount of revenue that a customer is expected to generate over their lifetime.
Monthly Recurring Revenue (MRR)	(Number of Paying Customers) * (Average Monthly Revenue per Customer)	The amount of recurring revenue that a business generates each month.
Annual Recurring Revenue (ARR)	MRR * 12	The amount of recurring revenue that a business generates each year.
Revenue Growth Rate	((Current Revenue - Previous Revenue) / Previous Revenue) * 100	The rate at which a company's revenue is growing.
Market Size	(Total Number of Potential Customers) * (Average Revenue per Customer)	The total size of the market for a particular product or service.
Market Share	(Company Revenue / Total Market Size) * 100	The percentage of the market that a company controls.
Revenue Potential	Market Size * Market Share	The amount of revenue that a company could potentially generate.

Optimizing marketing and sales efforts:

4) Customer Acquisition Cost (CAC): The cost of acquiring a new customer.

- LTV should be greater than CAC for a business to be profitable

- CAC: Can be calculated by dividing marketing and sales expenses by the number of new customers acquired.

- CAC helps businesses evaluate the effectiveness of their marketing and sales campaigns.

Optimizing marketing and sales efforts:

Metric	Formula	Description
Customer Acquisition Cost (CAC)	(Total Marketing and Sales Expenses) / (Number of New Customers Acquired)	The cost of acquiring a new customer.
Conversion Rate	(Number of Conversions) / (Number of Visitors)	The percentage of visitors to a website or landing page who take a desired action, such as making a purchase or signing up for a newsletter.
Sales Cycle Length	(Average Time to Close a Sale) / (Number of Sales)	The average amount of time it takes to close a sale.
Marketing Return on Investment (MROI)	(Revenue Generated from Marketing) / (Marketing Investment)	The amount of revenue that is generated for every dollar invested in marketing.
Sales Return on Investment (SROI)	(Revenue Generated from Sales) / (Sales Investment)	The amount of revenue that is generated for every dollar invested in sales.

Reducing customer churn:

5) Churn Rate: The percentage of customers who stop using a product or service over a given period of time.

- A high churn rate can lead to a decline in MAU and DAU.

- Churn Rate: Can be calculated by tracking the number of customers who cancel their subscriptions or stop using a service over a given period of time.

- Churn rate analysis helps businesses identify the reasons why customers are leaving and develop strategies to reduce churn.

Reducing customer churn:

Metric	Formula	Description
Customer Churn Rate	(Number of Customers Lost) / (Total Number of Customers)	The rate at which customers stop doing business with a company.
Customer Satisfaction Score (CSAT)	(Number of Satisfied Customers) / (Total Number of Customers)	The percentage of customers who are satisfied with a company's products or services.
Net Promoter Score (NPS)	(Number of Promoters - Number of Detractors) / (Total Number of Customers)	The percentage of customers who are likely to recommend a company's products or services to others.
Customer Effort Score (CES)	(Average Customer Effort Score) / (Total Number of Customers)	The average amount of effort that customers have to put in to do business with a company.

Key Business Metrics

Key Business Metrics for Movie Hosting:

- The growth and engagement of a movie hosting platform:

Total Views: The total number of times that movies have been viewed on the platform.

Average Views per User: The average number of times that a user views a movie on the platform.

Total Watch Time: The total amount of time that users have spent watching movies on the platform.

Average Watch Time per User: The average amount of time that a user spends watching movies on the platform.

- Understanding the profitability of a movie hosting platform

Content Acquisition Costs: The cost of acquiring new movies and TV shows for the platform.

Marketing and Advertising Costs: The cost of marketing and advertising the platform to potential users.

Customer Support Costs: The cost of providing customer support to users.

Infrastructure Costs: The cost of maintaining the platform's infrastructure, such as servers and bandwidth.

Lecture 15

Ksenia
Kasianova

Intro

Content
analytics

Web
Content
Monetiza-
tion

Search and
promotion
of content
in the
product

Technical
structure
of the
movie
hosting
website

Key
Business
Metrics

Dendrograms

End-to-
End
Analytics

Clustering
methods:

Key Business Metrics

Several Dashboards for Movie Hosting Content Analytics

1. Overview Dashboard

Metrics:

Monthly Active Users (MAU), Daily Active Users (DAU), Weekly Active Users (WAU)

Total Views, Average Views per User

Total Watch Time, Average Watch Time per User

Customer Churn Rate

Visualizations:

- Line charts showing the trends of these metrics over time

- Bar charts comparing the performance of different content types

- Pie charts showing the distribution of users by demographics

2. Engagement Dashboard

Metrics:

Average time spent watching movies

Number of movies watched per user

Number of likes, shares, and comments on movies

Number of user-created playlists

Visualizations:

- Line charts showing the trends of these metrics over time

- Bar charts comparing the performance of different content types

- Scatter plots showing the relationship between different engagement metrics

3. Content Performance Dashboard

Metrics:

Number of views for each movie

Average rating for each movie

Number of likes, shares, and comments for each movie

Customer churn rate for users who have watched a specific movie

Visualizations:

- Bar charts showing the top-performing movies

- Line charts showing the trends of these metrics over time

- Heat maps showing the distribution of views by geography

4. User Segmentation Dashboard

Metrics:

Demographics of users (age, gender, location, etc.)

Viewing habits of users (favorite genres, average watch time, etc.)

Churn rate of users in different segments

Visualizations:

- Pie charts showing the distribution of users by demographics

- Bar charts comparing the viewing habits of different user segments

- Line charts showing the churn rate of users in different segments

5. Marketing Dashboard

Metrics:

Cost per acquisition (CPA)

Return on investment (ROI) for marketing campaigns

Number of new users acquired through each marketing channel

Visualizations:

- Line charts showing the trends of these metrics over time

- Bar charts comparing the performance of different marketing channels

- Pie charts showing the distribution of new users by acquisition channel

In addition to the above dashboards, other useful dashboards for movie hosting content analytics include:

- Technical Dashboard: track the performance of the platform's infrastructure, such as server uptime and bandwidth usage.

- Financial Dashboard: track the platform's financial performance, such as revenue, expenses, and profit.

- Customer Support Dashboard: the performance of the platform's customer support team, such as the number of tickets resolved and the average response time.

End-to-end analytics is a process that encompasses all of the steps, from data collection to data-driven decision making.

Example of End-to-End Analytics

1. Data Collection: The platform could collect data from a variety of sources, such as its website, mobile app, and social media channels, including information about user demographics, viewing habits, and preferences.

- Data sampling, Data anonymization, Data encryption

2. Data Integration: integrate this data into a single, unified dataset.

- Schema mapping, Data merging, Data federation

3. Data Cleaning: removing errors and inconsistencies.

- Range checking, Type checking, Format checking, Checksums, Hashing

- Duplicates, missing values, outliers, or other anomalies.

4. Data Transformation: transform the data into a format that is suitable for analysis.

- Normalization, Standardization, Aggregation, Data integration

5. Data Analysis: extracting insights.

6. Data Visualization: identify trends and patterns in the data.

7. Data-Driven Decision Making: personalizing the user experience, recommend movies to users, and target marketing campaigns more effectively.