Extracting Features of Entertainment Products: A Guided Latent Dirichlet Allocation Approach Informed by the Psychology of Media Consumption

2018 Journal of Marketing Research

Olivier Toubia, Garud Iyengar, Renée Bunnell, and Alain Lemaire

汇报人：钱洋

Motivation

Olivier Toubia

关注

Glaubinger Professor of Business, Columbia Business School
在 columbia.edu 的电子邮件经过验证 - 首页

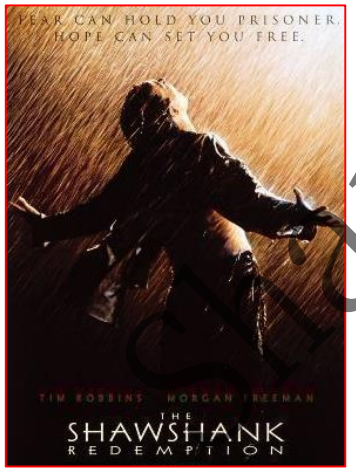
Marketing

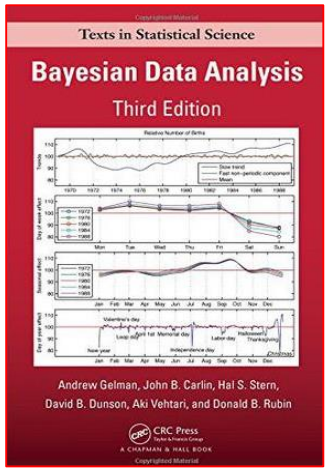
标题	引用次数	年份
A Poisson Factorization Topic Model for the Study of Creative Documents (and Their Summaries) O Toubia Available at SSRN 3334028		2019
Extracting Features of Entertainment Products: A Guided Latent Dirichlet Allocation Approach Informed by the Psychology of Media Consumption O Toubia, G Iyengar, R Bunnell, A Lemaire Journal of Marketing Research, 0022243718820559		2018
Attention, Information Processing, and Choice in Incentive-Aligned Choice Experiments L Yang, O Toubia, MG de Jong Journal of Marketing Research 55 (6), 783-800	3	2018
A semantic approach for estimating consumer content preferences from online search queries J Liu, O Toubia Marketing Science 37 (6), 930-952	4	2018
What's the Catch? Suspicion of Bank Motives and Sluggish Refinancing EJ Johnson, S Meier, O Toubia		2018

Puranam D, Narayan V, Kadiyali V. The effect of calorie posting regulation on consumer opinion: a flexible Latent Dirichlet Allocation model with informative priors[J]. Marketing Science, 2017, 36(5): 726-746.

Motivation

- **Propose a quantitative approach for describing entertainment products(娱乐产品)**
  - **Use guided latent Dirichlet allocation to extract a set of features of entertainment products from their descriptions (seed words).**
  - **People's consumption of entertainment products is influenced by the psychological themes.**
  - **Classify psychological themes on the basis of the "character strengths"(个性优势) taxonomy from the positive psychology literature.**
- **Improve the predictive performance of consumer choice models(predict movie-watching behavior at the individual level)**





Motivation

Input Data: Movie's synopses (IMDb)

Synopsis

In 2014/2015, bio-medical engineer Dr. Ryan Stone (Sandra Bullock) is a Mission Specialist on her first space shuttle mission, STS-157. She is on a spacewalk repairing a panel on the Hubble Space Telescope, accompanied by veteran astronaut Matt Kowalski (George Clooney), who is commanding his final expedition. Mission Control (voice of Ed Harris) suddenly instructs them to abort their spacewalk and return to the STS. Houston tells them that debris from a Russian missile strike on a defunct satellite has caused a chain reaction, destroying other satellites, and a huge debris field is heading right at them. Within a couple of minutes, they lose contact with Mission Control, though Stone and Kowalski continue to transmit "in the blind", hoping that the ground crew can hear them.

Extract features (guided LDA)

特征是有含义的，不是完全隐的

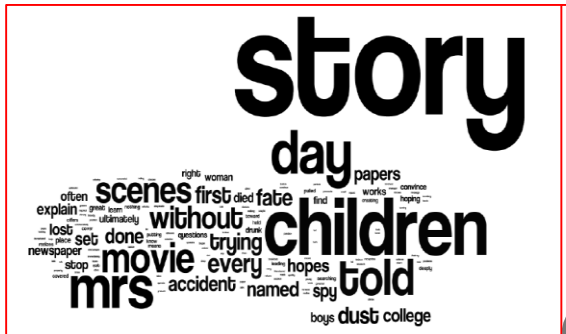
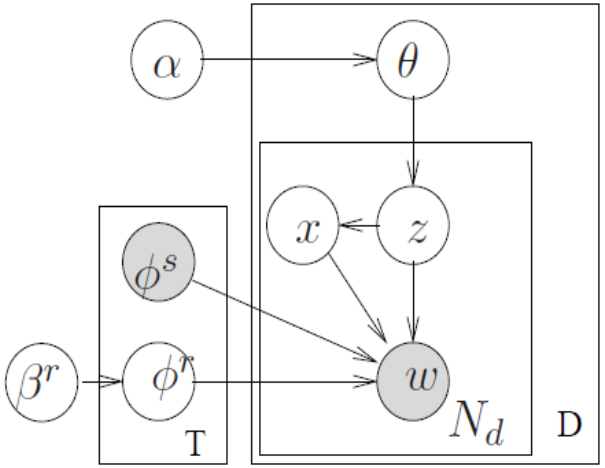


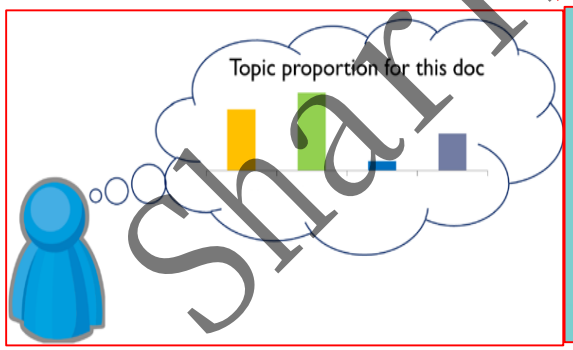
Figure WApp 2: "Creativity 4"

Figure WApp 3: "Fairness 1"



Interactive topic modeling (positive psychology)

Output and Following Work



Predict movie-watching behavior (content-based/hybrid approaches)

Introduction

- The revenue of the global entertainment and media industry was estimated at $1.8 trillion(1.8万亿) in 2016 (Statista 2017).
- One important trend in this industry is the increasing use of digital services(数字化服务) such as streaming(流媒体), video on demand(视频点播), e-readers(电子图书), and so on.
- Importantly for marketers, these technologies increase the availability of panel data.

How to leverage panel data in the media and entertainment industry?

Approaches

Introduction

◆ **Three categories:**

- **Pure collaborative approaches: a user's behavior is predicted on the basis of past behavior of similar users (**neighborhood-based collaborative filtering**)**
- **Content-based approaches: a user's behavior is predicted on the basis of his or her own past behavior (**regressions/ decision trees/neural networks to link product features to consumption**)**
- **Hybrid methods: combine collaborative and content-based methods(**content-boosted CF/ Bayesian Approach**)**

Introduction

◆ **In the marketing literature, most consumer choice models for entertainment products have been content-based or hybrid(作者列举了支撑文献).**

e.g., Ansari, Essegaier, and Kohli 2000; Bodapati 2008; Eliashberg and Sawhney 1994; Rust and Alpert 1984; Shachar and Emerson 2000; Ying, Feinberg, and Wedel 2006

- **Rely on estimating a set of weights on a preexisting set of product features(产品特征非常重要).**
- **Defining features for digital camera(功能型产品) is easy. When it comes to entertainment products, defining a feature set is not as straightforward.**

Introduction

genres

Introduction

◆ **Limitations of genre classifications:**

- **First, genres tend to be category-specific(e.g., “action,” “comedy”); not necessarily completely relevant in other industries (e.g., book).**

- **Second, and perhaps more importantly, traditional genre classifications are not enough to describe entertainment products with adequate granularity and richness(粗粒度).**

◆ **Limitations of collaborative filtering approaches：** (作者使用另外两种方法)

- **First, it becomes challenging to develop insights and reach interpretable results in the absence of a set of features that predict consumer choices(洞察力和可解释性).**

- **Second, collaborative approaches suffer from the “new item cold-start” problem.**

Introduction

◆ **Strengths of content-based and hybrid approaches：**

- **Allow predictions to be made for new products on the basis of consumers' preferences for the features that describe the content of these products.(消费者对产品特征的偏好)**

Introduction

◆ **Contributions:**

- **new and more powerful methods vs. better input for content-based and hybrid methods(别人做方法，我做输入)**

- **Our objective is not to develop new methods that predict consumer choices conditional on a set of features but rather to develop a new method for constructing the set of features, which can be used as input into any existing content-based or hybrid model that attempts to predict the behavior of consumers on the basis of past behavior. (再次强调)**

- **Our taxonomy is inspired by the psychology behind the consumption of entertainment products.(the positive psychology literature)**
- **Our guided LDA method is automated and scalable, guided LDA features may be used as input into any existing content-based or hybrid "big data" analytics tools, including the ones developed in the marketing literature.(再次强调)**

Literature

◆ Media Psychology(媒体心理学)

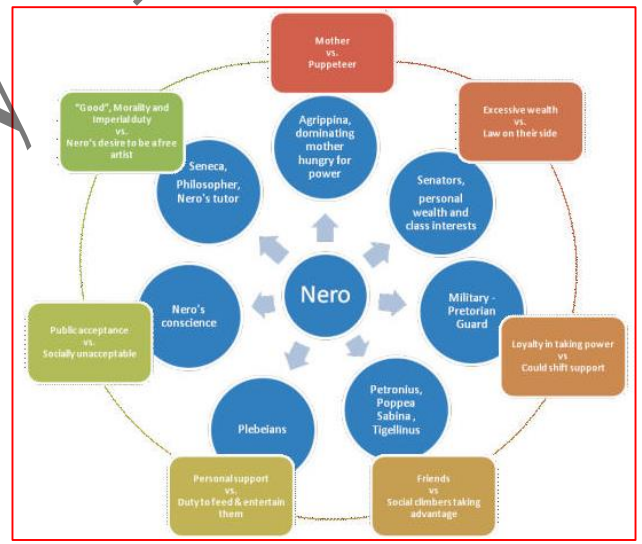
- 媒体心理学研究如何描述，解释，预测人们使用媒体或者与媒体相关的体验与行为(回应/互动等).

- **People prefer entertainment products that satisfy psychological needs.**

Consumer's psychological profile

Matching

Preferences



Psychological themes featured in the product

Literature

◆ **Media Psychology(媒体心理学)**

- **The approach of measuring consumers' psychological profiles and linking them to their media preferences has been useful in demonstrating that psychological factors are important predictors of media preferences.----- traditional approach: survey**

- **In contrast, we focus on describing the entertainment products themselves on the basis of the psychological themes they feature rather than describing consumers on the basis of their own psychological profiles.**

- **Producing features to be incorporated into models that learn consumers' preferences through their behavior.**

How to classify the psychological themes?

Literature

◆ **Positive Psychology(积极心理学/正向/正面心理学)**

- **Big Five framework: 5 dimensions (coarse-grained)**

- **Positive psychology framework: 24 dimensions (Fine-grained)**

Peterson C, Seligman M E P. Character strengths and virtues: A handbook and classification[M]. Oxford University Press, 2004. (7619)

Literature

◆ **Positive Psychology(积极心理学/正向/正面心理学)**

- **Clinical psychologists have previously attempted to establish connections between the positive psychology literature and the media and entertainment literature.**

Qualitative/descriptive

Vs.

Quantitative

- **We focus on the consumption of entertainment products, and we propose a scalable tool for automatically classifying products, without relying on human input.**

Literature

◆ **Screenwriting (剧本创作)**

- **The screenwriting literature has identified factors that describe movies and influence the quality of a script.**

- **Eliashberg, Hui, and Zhang(2007,2014) integrate and summarize this literature to construct a set of criteria that capture "how a story should be told and what kind of stories would resonate with audience".（如何描述一个故事，能够引起观众的共鸣） For example, the story should follow a logical, causal relationship, each scene description should advance the plot and be closely connected to the central conflict(因果关系、连贯性、剧情是否冲突).**

- **Our primary focus in this article is on predicting individual-level behavior captured by panel data. Accordingly, the features we extract from entertainment products are meant to reflect "horizontal" rather than "vertical" differentiation.（consumer tastes rather than differences in the overall quality of a story）**

Literature

◆ Natural Language Processing

- **We have hypothesized that the consumption of entertainment products may be linked to the psychological themes featured in these products, and that the positive psychology literature provides a useful taxonomy of psychological themes.(娱乐产品的消费与产品表达的心理学特征是密切联系的)**

- **Accordingly, we develop an approach that is flexible enough to allow features to be informed and guided by our taxonomy of psychological themes, while allowing other relevant dimensions to emerge.**

Literature

◆ Natural Language Processing

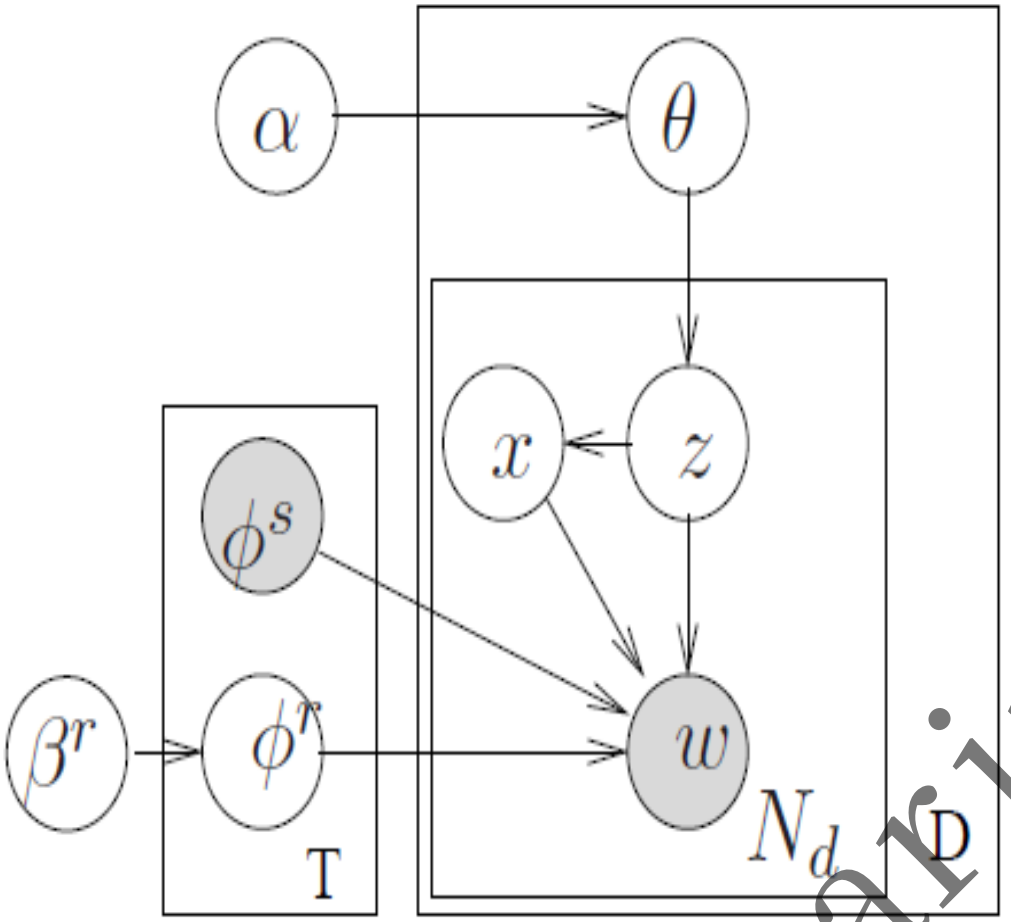
- **LDA and LSA in marketing research.**

- **In traditional LDA or in LSA, topics emerge strictly from the data and need to be labeled by the researcher. In our context, topics should be informed by psychological themes(主题应以心理学内容为依据).**

- **One approach would be to constrain each topic to reflect exactly one psychological theme by constraining the vocabulary in each topic to consist of a set of words that are known to be associated with a particular psychological theme.**

Guided LDA

◆ Model



1. For each topic k = 1, ..., K,
   - Draw regular topic: ϕ_k^r: Dirichlet($\alpha_1 l_k^r$)
   - Draw seed topic: ϕ_k^s: Dirichlet($\alpha_1 l_k^s$)
   - Draw weight on seeded topic: π_k: Beta(1, 1)
2. For each document d = 1, ..., D,
   - Draw topic distribution: θ_d: Dirichlet($\alpha_2 1_K$)
   - For each token i:
     - Draw a topic: z_i^d: Multinomial(θ_d)
     - Draw an indicator: x_i^d: Binomial($\pi_{z_i^d}$)
     - If indicator $x_i = 0$, draw a word from regular topic: w_i^d: Multinomial($\phi_{z_i^d}^r$)
     - If indicator $x_i = 1$, draw a word from seeded topic: w_i^d: Multinomial($\phi_{z_i^d}^s$)

Guided LDA

◆ Compiling the Set of Seed Words

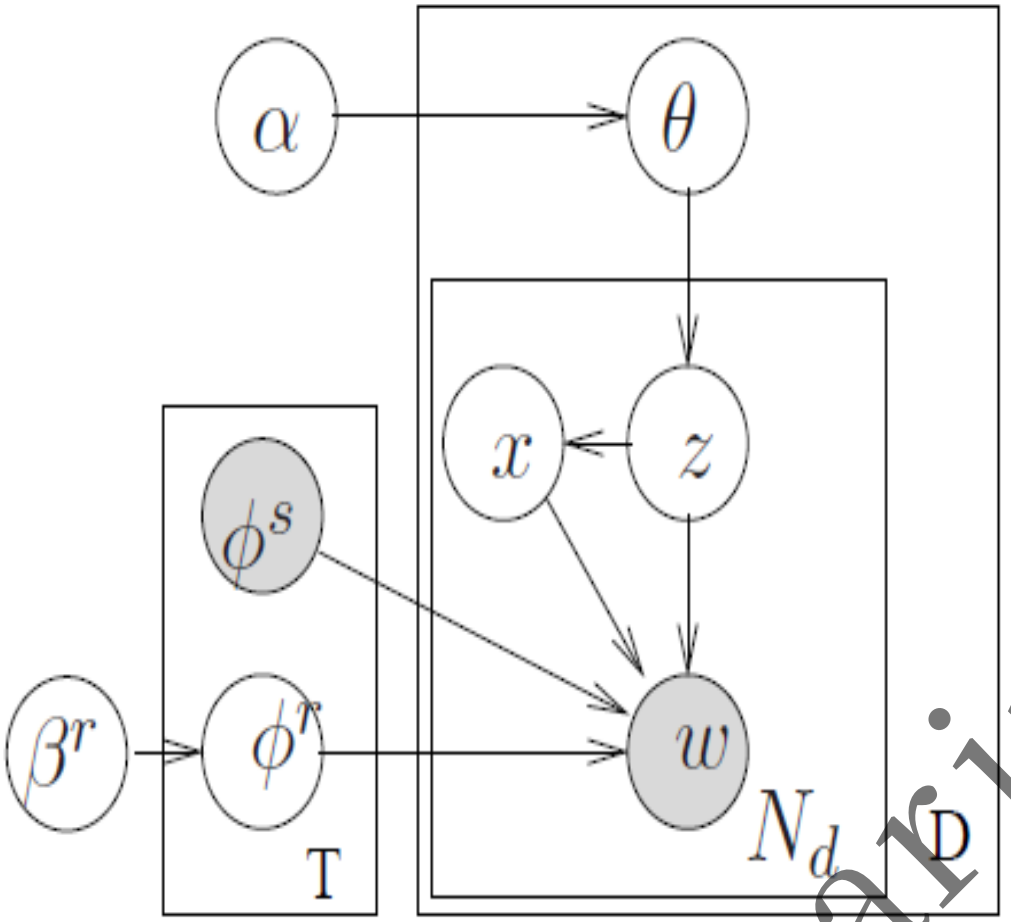
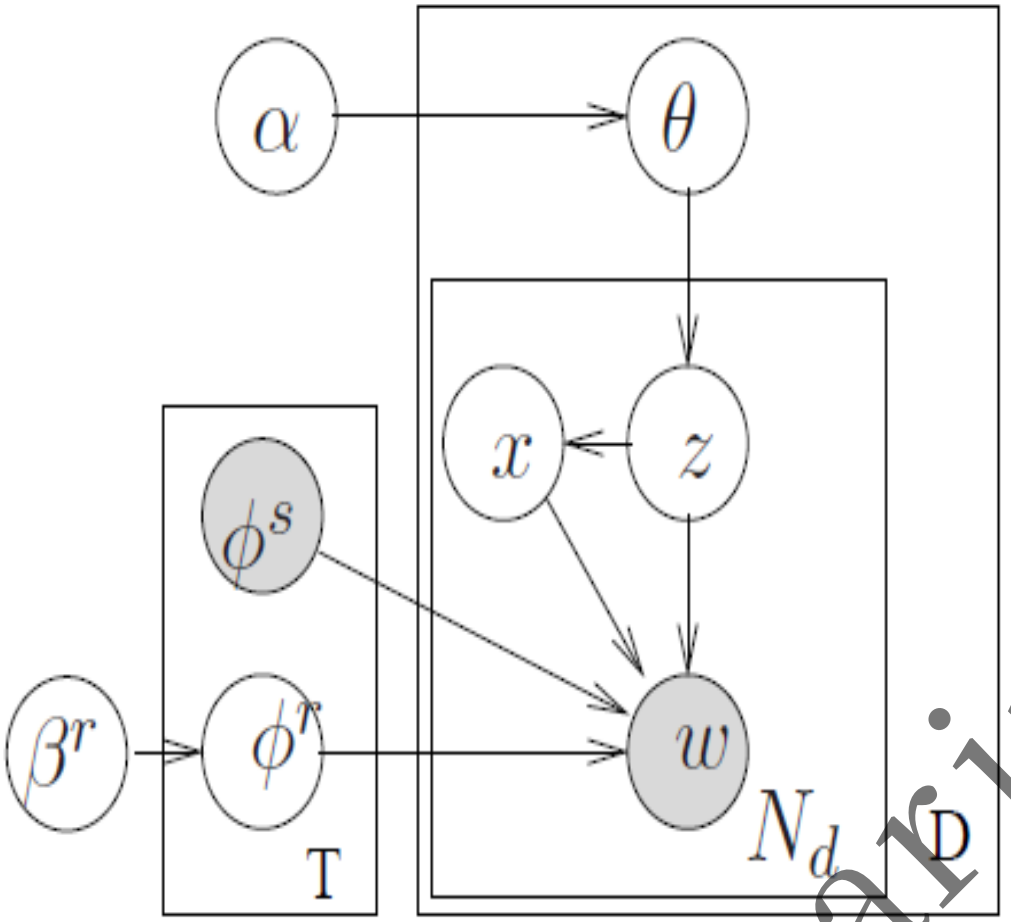


Table 1. List of Psychological Themes and Examples of Seed Words.

Psychological Theme	Examples of Seed Words
Creativity	idea, original, novel
Curiosity	discover, question, interested
Open Mindedness	examine, considerate, impartial
Love of Learning	school, course, professor
Wisdom	experience, knowledge, advisor
Bravery	battle, hero, courage
Persistence	goal, effort, sacrifice
Integrity	truth, promise, genuine
Vitality	energy, peppy, enthusiastic
Love	relationship, marriage, friend
Kindness	gift, favor, compassion
Social intelligence	psychologist, mindful, insightful
Citizenship	loyal, society, duty
Fairness	justice, law, rule
Leadership	team, captain, president
Forgiveness and mercy	apologize, peace, repent
Humility and modesty	humble, discrete, timid
Prudence	careful, responsible, safety
Self-regulation	abstain, restrain, virgin
Appreciation of beauty and excellence	wonder, awe, beautiful
Gratitude	gift, grateful, blessed
Hope	dream, opportunity, confidence
Humor	joke, laugh, funny
Spirituality	church, faith, heaven

Guided LDA

◆ Compiling the Set of Seed Words



We selected ten common seed words for each of the 24 psychological themes, based on our preliminary analysis.【根据作者的初步分析】

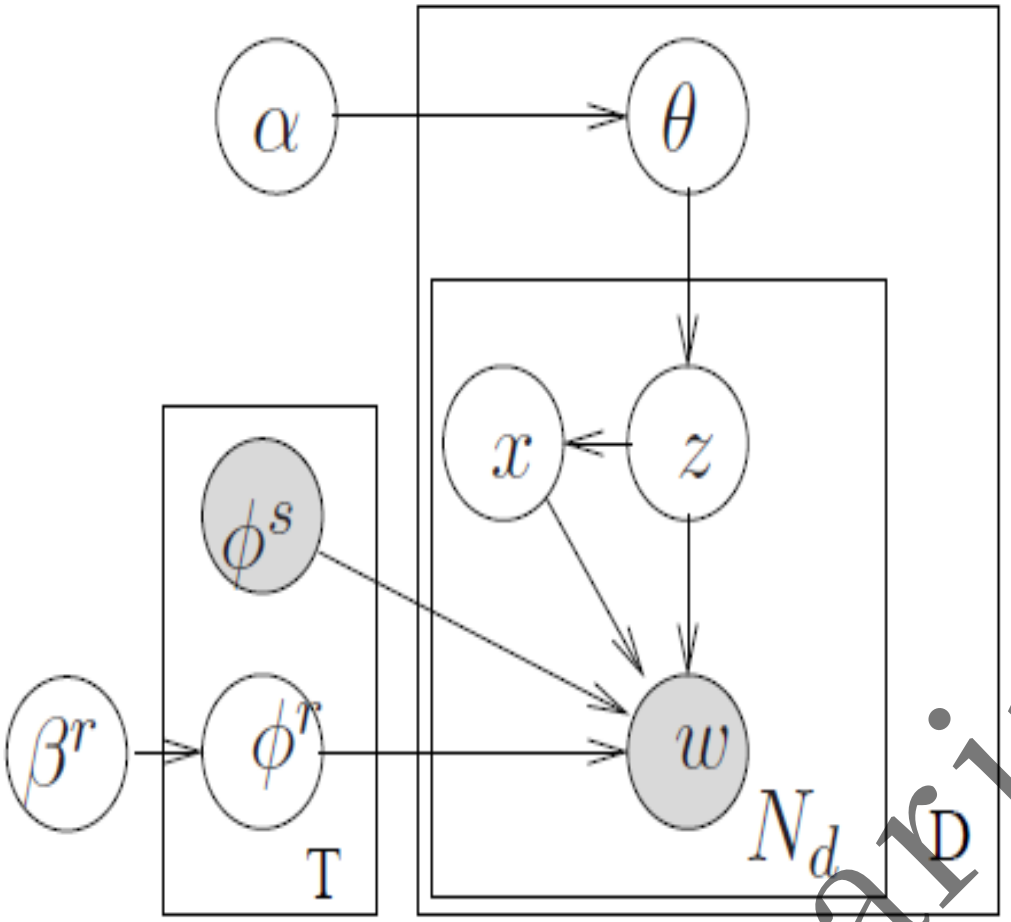
For each theme, we showed participants the ten seed words and asked them to propose three new words that would complement the list well.

from N =106 respondents(paid $1)

Our final dictionary of seed words contains 2,677 unique seed words.

Guided LDA

◆ Creating the Vocabulary



Term frequency–inverse document frequency

We keep the 2,000 words with the highest tf–idf.

2,677+2,000=4,677

$$w = 1, \ldots, W$$

W is the "all other" word.

$$K = 24n + 1$$ → subthemes

The last topic, K, is a baseline topic. (this topic may have positive weights on all words, including the "all other" word)

Application to Movies

◆ Movie Descriptions

- **We use movie synopses available on IMDb as input into our guided LDA analysis.**
- **Benefits of synopses： First, they are not unique to the movie industry, and they are available for most entertainment products. Second, compared with reviews, synopses have the benefit of being objective descriptions rather than subjective evaluations.**
- **We assembled a data set of 429 movie descriptions. (**39 movies that received one of the “big five” Oscars, and the top 40 movies, in terms of U.S. domestic box office performance**):** Study 1 and Study 2
- **Preprocessing(**R tm package**): non-English characters and words and tokenized the text.**
- **For robustness, we repeated the analysis with two other data sources: movie spoilers and scripts.(电影剧透和电影剧本)**

Application to Movies

◆ Movie Descriptions

Table 4. Descriptive Statistics of Movie Descriptions (Synopses).

Statistic	Unit of Analysis	Mean	SD	Min	Max
Number of words (including "all other")	Movie descriptions (N = 429)	1,446.65	1,226.82	42	5,817
Number of occurrences of seed words	Movie descriptions (N = 429)	72.19	55.73	3	397
Number of unique seed words	Movie descriptions (N = 429)	45.25	27.97	3	167
Number of psychological themes with at least one seed word occurrence	Movie descriptions (N = 429)	18.43	4.26	3	24
Total number of occurrences across movie descriptions	Seed words (N = 2,677)	11.57	36.77	0	624
Proportion of movie descriptions with at least one occurrence	Seed words (N = 2,677)	.02	.04	0	.61
Total number of occurrences across movie descriptions	Seed words with at least one occurrence (N = 1,608)	19.26	45.86	1	624
Proportion of movie descriptions with at least one occurrence	Seed words with at least one occurrence (N = 1,608)	.03	.05	.002	.61
Average number of seed word occurrences per movie description	Psychological theme (N = 24)	4.03	2.71	1.25	13.08
Proportion of movie descriptions with at least one seed word occurrence	Psychological theme (N = 24)	.77	.12	.52	.97

Application to Movies

◆ Guided LDA Results

- **To inform model selection, we compute the deviance information criterion (DIC)**

Table 5. Guided LDA Versus Traditional LDA.

Number of Topics per Psychological Theme (n)	Total number of Topics	DIC for Guided LDA ($\times 10^3$)	DIC for Traditional LDA ($\times 10^3$)
1	25	2,043.9	2,073.3
2	49	1,781.0	1,819.4
3	73	1,659.3	1,697.2
4	97	1,554.9	1,594.9

As we have noted, increasing n further led to convergence issues when estimating choice models on the data collected in Studies 1 and 2.

$$(\phi_k(w) = \pi_k \phi_k^s(w) + (1 - \pi_k)\phi_k^r(w)),$$

Application to Movies

◆ Guided LDA Results

Table 6. Examples of Topics from Guided LDA.

Topic	Average Document-Topic Weight ($\times 10^{-3}$)	Example of Movie with Large Weight	Examples of Words with High Relevance Present in Movie Description
"Citizenship 4"	4.07	*My Big Fat Greek Wedding*	family, daughter, time
"Creativity 4"	4.30	*The Golden Compass*	children, told, dust
"Fairness 1"	3.47	*Robin Hood*	king, sword, lady
"Leadership 3"	5.02	*Glory Road*	team, coach, players
"Leadership 4"	3.33	*G.I. Joe: Retaliation*	president, storm, shadow
"Love 1"	4.24	*The Secret Life of Bees*	mother, growing, bed
"Love 3"	3.47	*Kissing Jessica Stein*	friend, night, girl
"Love 4"	5.41	*Sex and the City*	wedding, marriage, affair
"Love of Learning 3"	4.61	*Freedom Writers*	students, school, class
"Vitality 3"	4.80	*Eat, Pray, Love*	life, returns, experience

◆ Empirical Framework

- We continue with our application to movies, focusing on individual-level consumption. That is, our dependent variable is whether a particular consumer chose to watch a particular movie.

- We consider data that capture binary viewing decisions made by C consumers on M movies.

- We specify a simple predictive content-based model that links product features to movie consumption. (binomial logistic choice probabilities)

$$\text{Prob}(y_{cm} = 1) = \frac{\exp(X_m W_c)}{1 + \exp(X_m W_c)}$$

◆ Design of the Studies (Study 1)

- **Study 1 focused on movies that won one of the "big five" Oscars between 2004 and 2014 (39).**

- **We recruited participants from MTurk's online panel, screened for being based in the United States. We asked each respondent to indicate whether (s)he had watched each of the movies in the set. We received complete data from N=599 participants, who were each paid $1 for their participation.**

- **Each movie had been watched by an average of 33.19% of the participants (SD=16.57%), and each participant had watched, on average, 12.94 of the movies in the sample (SD=7.84).**

◆ Design of the Studies (Study 2)

- **We selected the top 40 movies based on U.S. domestic box office performance in 2013.**

- **We recruited participants from MTurk's online panel, screened for being based in theUnited States. Again, we asked each respondent to indicate whether (s)he had watched each of the movies in the set. We received complete data from N=542 respondents, who were each paid $1 for their participation.**

- **Each movie had been watched by an average of 30.33% of the participants (SD=11.11%), and each participant had watched, on average, 12.13 of the movies in the sample (SD = 8.30).**

◆ Movie Features

- We consider three sets of predictive variables (features) that may be used to describe movies and predict this dependent variable at the consumer level.

(1) The first set of features capture information about movies that is commonly considered in academic studies on movies.[average critic rating; the average user score; the production budget; widest release(电影院放映的场次); the domestic box office performance(票房); the Motion Picture Association of America (MPAA) rating(美国电影协会评级); the movie's run time in minutes(电影的时长分钟); a dummy variable equal to 1 if the movie was a sequel(电影是否为续集); the degree of competition faced by the movie at the time of its release(电影发行时，面临的竞争); star power(明星的影响STARmeter rating provided by IMDb); Twitter activity(有关该电影的推特数量); DVD release timing; DVD sales rank]

这些变量都是有依据的，来源于不同网站的

◆ Movie Features

- We consider three sets of predictive variables (features) that may be used to describe movies and predict this dependent variable at the consumer level.

(2) The second set of features capture the content of each movie and are based on the work of Eliashberg, Hui, and Zhang (2007,2014). Genres and Content variables(read the script of each movie/questionnaire Eliashberg,Hui, and Zhang); Semantic variables(spoilers(剧透): the number of characters, the number of words, the number of sentences, and the average number of characters per word); Bag-of-words variables from LSA(following Eliashberg, Hui, and Zhang)

(3) The third and final set of features consists of the weights estimated by guided LDA, capturing the extent to which movie features each topic.

这些变量都是有依据的，来源于不同网站的

◆ Movie Features

Table 7. Variables in Studies 1 and 2.

Variables	Type	Description	Source
Movie watching	Dependent	Dummy variable $y_{cm} = 1$ if consumer c watched movie m	Survey
Average critic rating	Predictive	Continuous variable between 0 and 100	Metacritic
Average user score	Predictive	Continuous variable between 0 and 10	Metacritic
Production budget (in $M)	Predictive	Continuous variable (inflation adjusted)	IMDb
Widest release (in thousands of theatres)	Predictive	Continuous variable	Box Office Mojo
Widest release (in thousands of theatres)^2	Predictive	Continuous variable	Box Office Mojo
Domestic box office (in $M)	Predictive	Continuous variable (inflation adjusted)	IMDb
MPAA rating	Predictive	Dummy variable(s)	IMDb
Run time (in minutes)	Predictive	Count variable	Box Office Mojo
Sequel	Predictive	Dummy variable	IMDb
Competition	Predictive	2 dummy variables	IMDb
Star power	Predictive	Discrete variable	IMDb
Twitter activity	Predictive	Discrete variable	MovieTweetings
DVD release timing	Predictive	Discrete variable (time elapsed between theatre and DVD release, in days)	IMDb and Amazon
DVD sales rank	Predictive	Discrete variable	Amazon
Genres	Predictive	8 variables	Independent raters (following Eliashberg, Hui, and Zhang [2014])
Content variables	Predictive	24 variables	Independent raters (following Eliashberg, Hui, and Zhang [2007])
Semantic variables	Predictive	4 variables	Word processing of spoilers (following Eliashberg, Hui, and Zhang [2007])
Bag-of-words variables from LSA	Predictive	2 continuous variables	LSA on spoilers (following Eliashberg, Hui, and Zhang [2007])
Guided LDA topic weights	Predictive	96 continuous variables between 0 and 1	Guided LDA

◆ Leveraging Guided LDA Features in Content-Based Choice Models

- **We use a hierarchical Bayes logistic choice model.**

$$\text{Prob}(y_{cm} = 1) = \frac{\exp(X_m W_c)}{1 + \exp(X_m W_c)}$$

$$W_c \sim N(W_0, D),$$

$$\Sigma \sim IW_{\nu_0}(\Lambda_0^{-1})$$

$$\mu|\Sigma \sim N(\mu_0, \Sigma/\kappa_0)$$

100,000 iterations, using the first 50,000 as burn-in and saving 1 in 10 iterations.

可参考：Conjugate Bayesian analysis of the Gaussian distribution

◆ Leveraging Guided LDA Features in Content-Based Choice Models

- **Our main focus in this section is on comparing the value of guided LDA features with that of other features based on the content. To that effect, we test whether guided LDA may complement or replace some of the features developed by Eliashberg, Hui, and Zhang.**

- **Version 1: We start with a specification of the choice model with an intercept only as a baseline (version 1).**

- **Version 2: We consider the inclusion of basic movie features.**

- **Version 3: We consider the addition of the features based on Eliashberg, Hui, and Zhang.**

◆ Leveraging Guided LDA Features in Content-Based Choice Models

- **Our main focus in this section is on comparing the value of guided LDA features with that of other features based on the content. To that effect, we test whether guided LDA may complement or replace some of the features developed by Eliashberg, Hui, and Zhang.**

- **Version 4: We consider replacing the bag-of-words variables created using LSA with guided LDA features.**

- **Version 5: We consider replacing all of the features based on Eliashberg, Hui, and Zhang (2007, 2014) with guided LDA features (version 5).**

◆ Leveraging Guided LDA Features in Content-Based Choice Models

Table 8. Study 1 Results: Pure Content-Based Choice Model.

Features	Version 1	Version 2	Version 3	Version 4	Version 5
Intercept	✓	✓	✓	✓	✓
Average critic rating		✓	✓	✓	✓
Average user score		✓	✓	✓	✓
Production budget		✓	✓	✓	✓
Widest release		✓	✓	✓	✓
Widest release^2		✓	✓	✓	✓
Domestic box office		✓	✓	✓	✓
MPAA rating		✓	✓	✓	✓
Run time		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star power		✓	✓	✓	✓
DVD release timing		✓	✓	✓	✓
DVD sales rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-words variables from LSA			✓		
Guided LDA topic weights				✓	✓
DIC	492.91	406.50	371.65	232.02	280.28
In-sample hit rate	62.09%	71.78%	76.30%	88.21%	85.08%
Out-of-sample hit rate	61.67%	66.44%	67.94%	70.32%	71.19%

Table 9. Study 2 Results: Pure Content-Based Choice Model.

Features	Version 1	Version 2	Version 3	Version 4	Version 5
Intercept	✓	✓	✓	✓	✓
Average critic rating		✓	✓	✓	✓
Average user score		✓	✓	✓	✓
Production budget		✓	✓	✓	✓
Widest release		✓	✓	✓	✓
Widest release^2		✓	✓	✓	✓
Domestic box office		✓	✓	✓	✓
MPAA rating		✓	✓	✓	✓
Run time		✓	✓	✓	✓
Sequel		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star power		✓	✓	✓	✓
Twitter activity		✓	✓	✓	✓
DVD release timing		✓	✓	✓	✓
DVD sales rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-words variables from LSA			✓		
Guided LDA topic weights				✓	✓
DIC	492.91	406.50	371.65	232.02	280.28
In-sample hit rate	64.05%	73.12%	76.54%	86.35%	83.28%
Out-of-sample hit rate	63.60%	68.91%	69.93%	71.00%	70.89%

◆ Leveraging Guided LDA Features in Hybrid Models: Content-Boosted CF

Melville, Prem, Raymond J. Mooney, and Ramadass Nagarajan (2002), "Content-Boosted Collaborative Filtering for Improved Recommendations," in Proceedings of the Eighteenth National Conference on Artificial Intelligence. New York: Association for Computing Machinery, 187–92.

Table 10. Study 1 Results: Content-Boosted Collaborative Filtering (CBCF).

Features	Pure CF	CBCF Version 2	CBCF Version 3	CBCF Version 4	CBCF Version 5
Intercept		✓	✓	✓	✓
Average critic rating		✓	✓	✓	✓
Average user score		✓	✓	✓	✓
Production budget		✓	✓	✓	✓
Widest release		✓	✓	✓	✓
Widest release^2		✓	✓	✓	✓
Domestic box office		✓	✓	✓	✓
MPAA rating		✓	✓	✓	✓
Run time		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star power		✓	✓	✓	✓
DVD release timing		✓	✓	✓	✓
DVD sales rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-words variables from LSA			✓		
Guided LDA topic weights					✓
Out-of-sample hit rate	68.67%	66.83%	68.05%	69.90%	70.60%

Table 11. Study 2 Results: Content-Boosted Collaborative Filtering (CBCF).

Features	Pure CF	CBCF Version 2	CBCF Version 3	CBCF Version 4	CBCF Version 5
Intercept		✓	✓	✓	✓
Average critic rating		✓	✓	✓	✓
Average user score		✓	✓	✓	✓
Production budget		✓	✓	✓	✓
Widest release		✓	✓	✓	✓
Widest release^2		✓	✓	✓	✓
Domestic box office		✓	✓	✓	✓
MPAA rating		✓	✓	✓	✓
Run time		✓	✓	✓	✓
Sequel		✓	✓	✓	✓
Competition		✓	✓	✓	✓
Star power		✓	✓	✓	✓
Twitter activity		✓	✓	✓	✓
DVD release timing		✓	✓	✓	✓
DVD sales rank		✓	✓	✓	✓
Genres			✓	✓	
Content variables			✓	✓	
Semantic variables			✓	✓	
Bag-of-words variables from LSA			✓		
Guided LDA topic weights				✓	✓
Out-of-sample hit rate	68.27%	68.66%	69.47%	70.31%	70.24%

◆ Using Guided LDA Features as Input into Predictive Models of Aggregate Performance

$$y_i = \log(\text{BOX OFFICE}_i / \text{BUDGET}_i)$$ 投资产出比

- We focus on the combined set of movies from Studies 1 and 2.
- We again use the genres, content, semantic, and bag-of-words variables (based on Eliashberg, Hui, and Zhang 2007, 2014).
- We test two models, based respectively on Eliashberg, Hui, and Zhang (2007) and Eliashberg, Hui, and Zhang (2014). a bootstrap aggregated classification and regression tree (bagged-CART) model based on the Bag-CART model; a kernel-based model based on the Kernel-II (optimized feature weights) model.
- We split the sample into 65 movies for calibration and 14 movies for validation. (replicating the analysis 100 times)

◆ Using Guided LDA Features as Input into Predictive Models of Aggregate Performance

Table 12. Using Guided LDA Features in Models that Predict Aggregate Performance.

Features	Bagged-CART			Kernel-Based		
Genres	✓	✓		✓	✓	
Content variables	✓	✓		✓	✓	
Semantic variables	✓	✓		✓	✓	
Bag-of-words variables from LSA	✓			✓		
Guided LDA topic weights		✓	✓		✓	✓
Out-of-sample MSE	.5186	.4590	.4657	.5176	.4703	.4806

Conclusions

- **We bridge the media psychology literature, the positive psychology literature, the NLP literature, the choice modeling literature, and the CF literature.**

- **We propose a new set of descriptors of entertainment products, theoretically founded in the media psychology literature and the positive psychology literature.**

- **We rely on the NLP literature to develop a method for tagging entertainment products in an automated and scalable manner.**

- **We first show that the proposed features improve our ability to predict consumption at the individual level.**

Summary

问题可以简单，但立意必须深远。

无论是方法驱动还是问题驱动，方法都要和问题完美融合。

写作要细节、表达要深刻。

谢谢