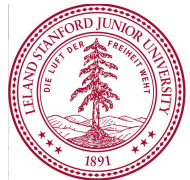


Why We Read Wikipedia

Philipp Singer*, Florian Lemmerich*, Robert West, Leila Zia,
Ellery Wulczyn, Markus Strohmaier, Jure Leskovec



Why readers?

- 610 million Wikipedia pageviews per day (October 30, 2016)
- 85% of these views came from users (*vs.* spiders and bots)

Blink your eyes once!

Why readers? (cont'd)

2000 Wikipedia pages were requested by humans as you were blinking. -- And we don't know why!

Providing educational content and effectively disseminating it requires understanding the needs and motivations of the people behind these pageviews.

Let's understand Wikipedia readers!

Literature

- Motivations and user behavior on the web (Goel et al. '12, Kumar and Tomkins '10), search engines (Broder '02, Rose and Levinson '04, Weber and Jaimes, '11), Twitter (Java et al. '07, Kwak et al. '10) and Facebook (Ryan and Xenos '11)
- Wikipedia editor motivations (Arazy et al. '17, Nov '07)
- Patterns of editing behavior (Jurgens and Lu, '12)
- Content preference (Lehmann '14, Ratkiewicz '10, Spoerri '07)
- Search queries leading to Wikipedia (Waller '11)
- Navigation patterns (Lamprecht et al. '16, Paranjape et al. '16, Singer et al. '14, West and Leskovec '12)
- New readers, https://meta.wikimedia.org/wiki/New_Readers

Contributions

1. A robust taxonomy for characterizing use cases for reading Wikipedia
2. Quantifying the prevalence and interactions between these use cases via a large-scale survey on English Wikipedia
3. Enhanced understanding of behavioral patterns associated with different use cases by combining survey responses with webrequest logs

A robust taxonomy of Wikipedia readers' use cases

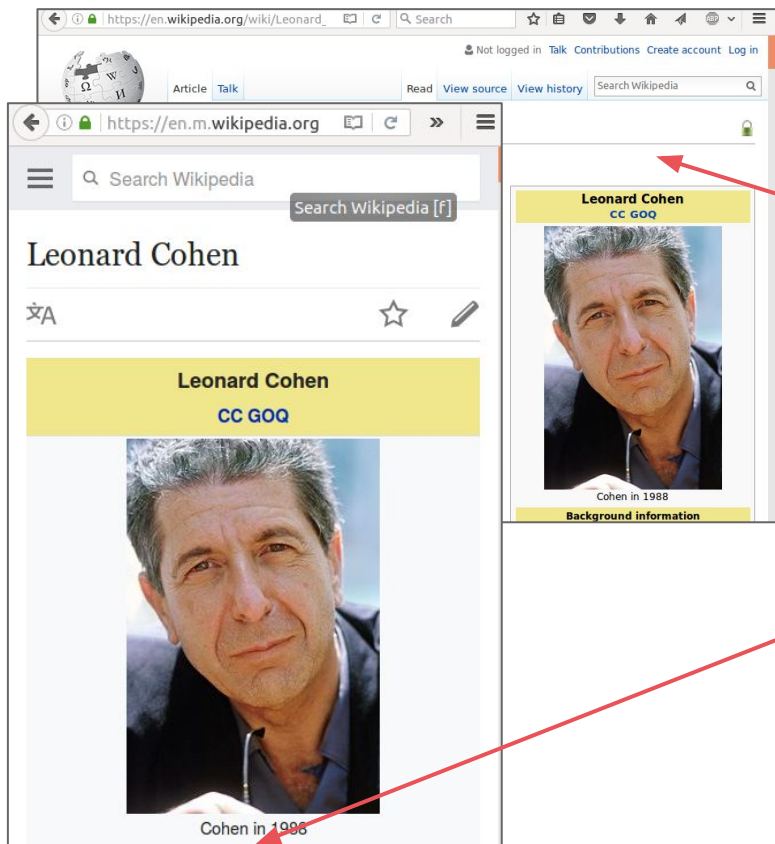
Where to start?

- Webrequest logs
 - Contains logs of all the hits to the WMF's servers
 - Each log includes a variety of information about the hit
 - Logs alone won't answer "why" readers come to Wikipedia
- Surveys
 - To understand why, we need to communicate with users at large scale

Survey 1: Building the initial taxonomy

- Duration: 4 days
- Sampling rate: 1 out of 200
- English Wikipedia, Mobile and Desktop
- On article pages
- Population: 5000 on Desktop, 1655 on Mobile.

Survey 1 - Widget



Answer one question and help us improve Wikipedia.

[Visit survey](#)

[No thanks](#)

Survey data handled by a third party. [Privacy](#)

* Required

Why are you reading this article today? *

There is a 100 character limit for your response.

[Submit](#)

Never submit passwords through Google Forms.

Powered by
 Google Forms

This form was created
inside of Wikimedia
Foundation.

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Survey 1 - Responses

“Personal interest about conflicts in middle east”

“So I can see the country’s population”

“Confirming address for shipment going to this town”

“NY Times today mentioned Operation Wetback, alluded to by Trump in debate, & wanted to learn more.”

“Studying for my med school test”

“Interest and curiosity”

“Because I am in a very boring art lesson”

“To find out more information about this aircraft.”

“Someone came by my desk talking about The Last Man on Earth

“To see a movie summary”

“I had previously edited it.”

(movie). So I looked it up.”

Survey 1 - Hand-coding

- Stage 1: went over 20 entries to build a common understanding.
- Stage 2: generously assigned tags to 500 randomly selected responses. Four main trends were identified.
- Stage 3: tagged 500 new responses.

For example: “To evaluate technical description of Bosch fuel injection system install on a car I’m interested in” -> tags: deep dive, shopping, technical. -> decision making, in-depth

Survey 1 - Output

- **Information need:** quick fact look-up, overview, in-depth reading.
- **Prior knowledge:** familiar, unfamiliar
- **Motivation:** work/school project, personal decision, current event, media, conversation, bored/random, intrinsic learning

Survey 2 & 3: Assessing the robustness

- **Survey 2:** are we missing categories applicable to other languages?
 - Repeated survey 1 in Persian and Spanish Wikipedia
- **Survey 3:** are we capturing all categories and dimensions?
 - Ran a 3-question survey in English Wikipedia with “Other” option for each question.
 - Only 2.3% of the responses chose “Other”.

Conclusion

We built a robust taxonomy of Wikipedia readers through a series of large scale surveys.

- **Information need:** fact look-up, overview, in-depth
- **Prior knowledge:** familiar, unfamiliar
- **Motivation:** work/school project, personal decision, current event, media, conversation, bored/random, intrinsic learning.

Quantifying the prevalence and interactions between use cases

Survey

- Duration: 1 week
- Sampling rate: 1 out of 50
- English Wikipedia, Mobile and Desktop
- On article pages and to those with DNT off.
- Population: 29,372

Why are you reading this article today?

I am reading this article to

- look up a specific fact or to get a quick answer.
- get an overview of the topic.
- get an in-depth understanding of the topic.

Prior to visiting this article

- I was not familiar with the topic and I am learning about it for the first time.
- I was already familiar with the topic.

I am reading this article because

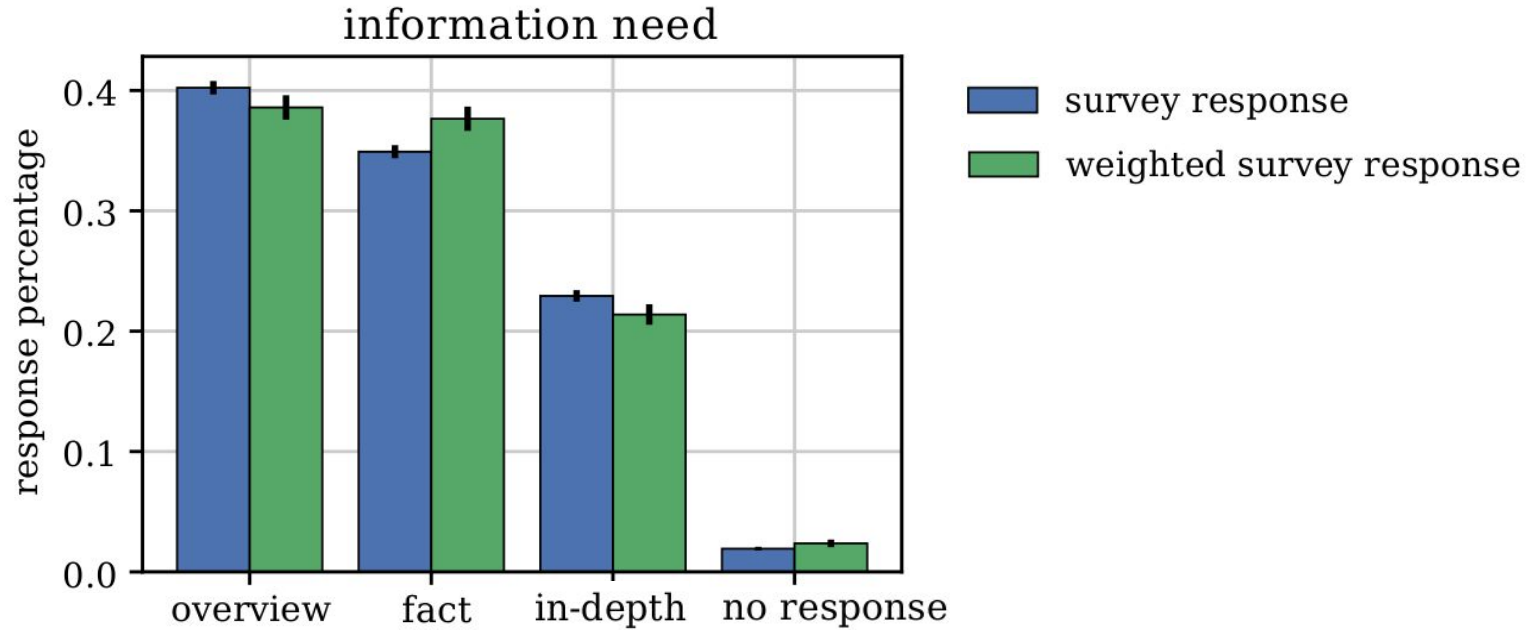
Please select all answers that apply

- the topic was referenced in a piece of media (e.g. TV, radio, article, film, book).
- I need to make a personal decision based on this topic (e.g. to buy a book, choose a travel destination).
- I am bored or randomly exploring Wikipedia for fun.
- the topic came up in a conversation.
- I have a work or school-related assignment.
- I want to know more about a current event (e.g. a soccer game, a recent earthquake, somebody's death).
- this topic is important to me and I want to learn more about it. (e.g., to learn about a culture).
- Other:

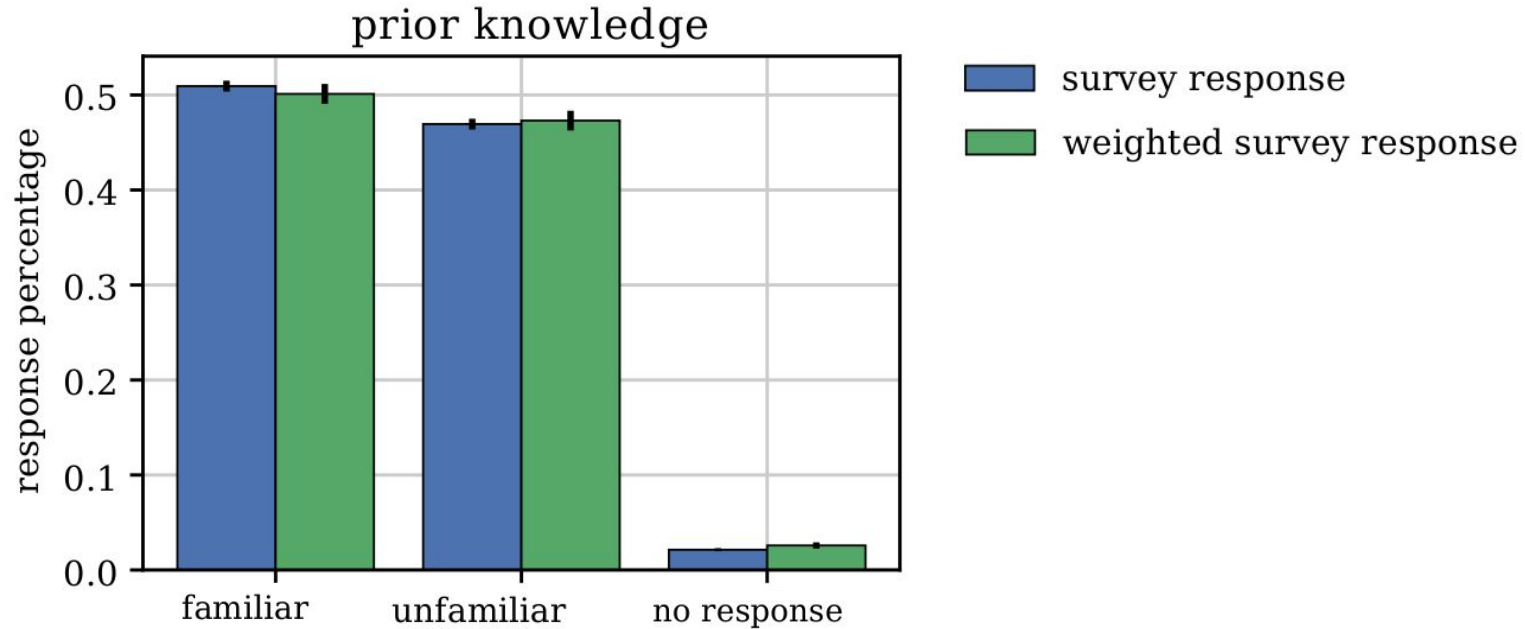
What about bias?

- Different kinds of bias may be at play, for example:
 - Those who had longer sessions were more likely to see the survey.
 - If you had a deadline for a project when the survey was shown to you, you might have been less likely to participate than if you were reading Wikipedia because you were bored.
 - ...
- Inverse propensity score weighting:
 - E.g., suppose a user subpopulation is represented two times more in the sample population when compared to the true population.
 - This can be accounted for by weighting the responses of this group by a factor of 0.5.

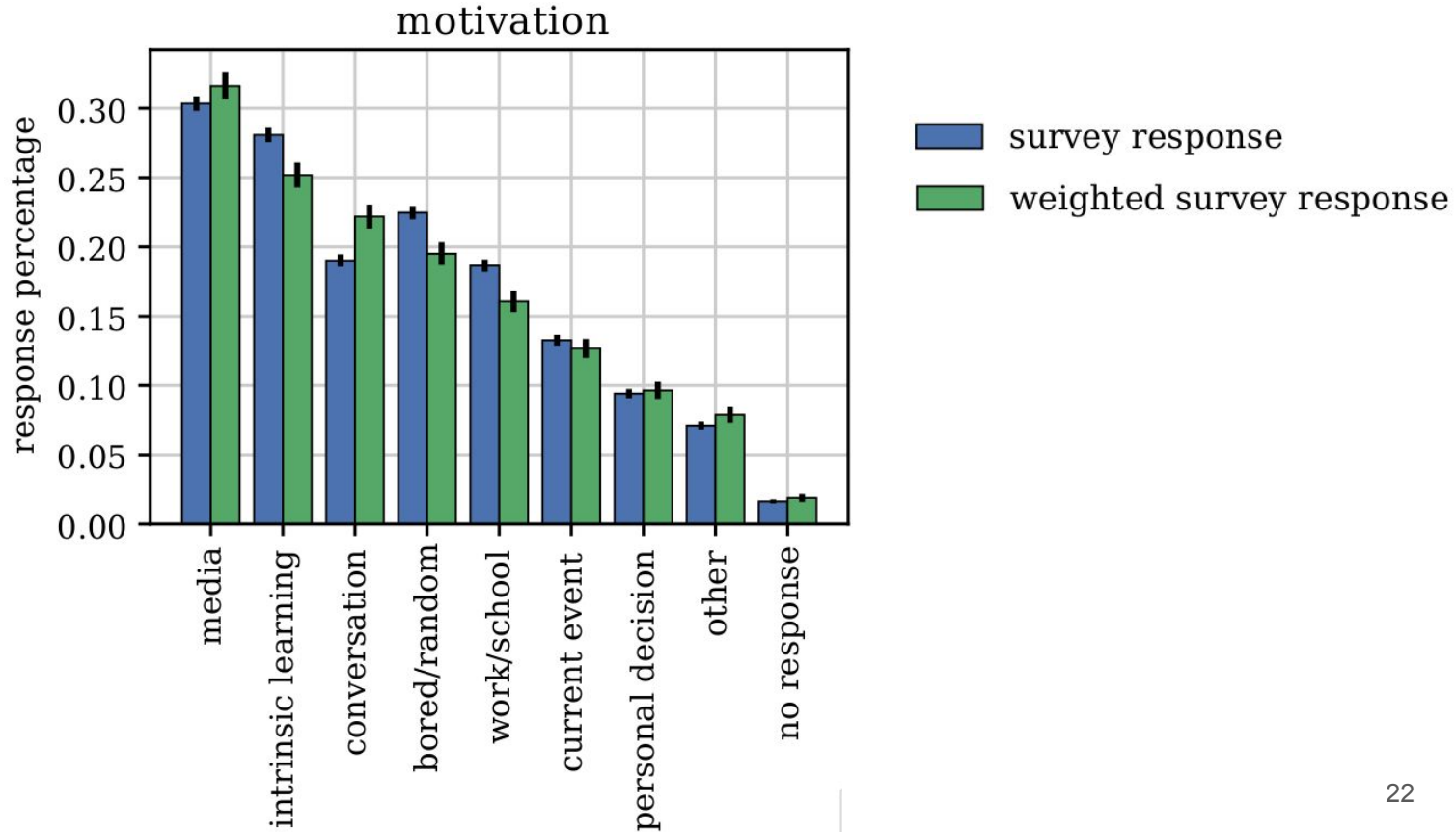
Information need



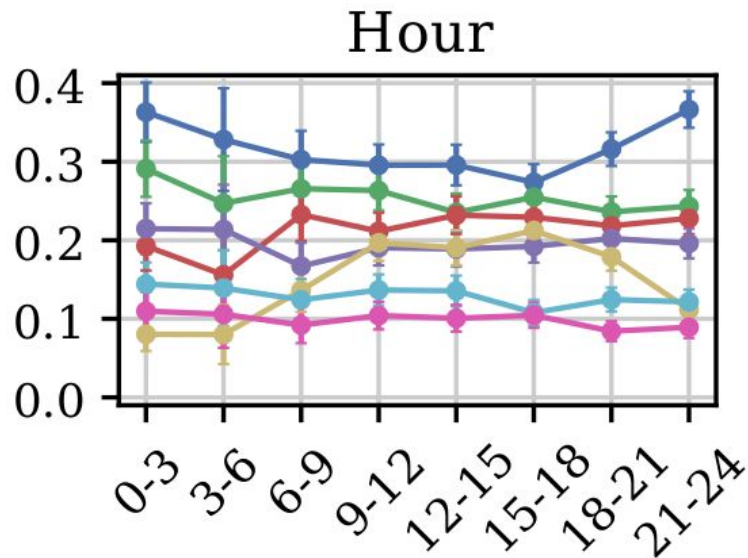
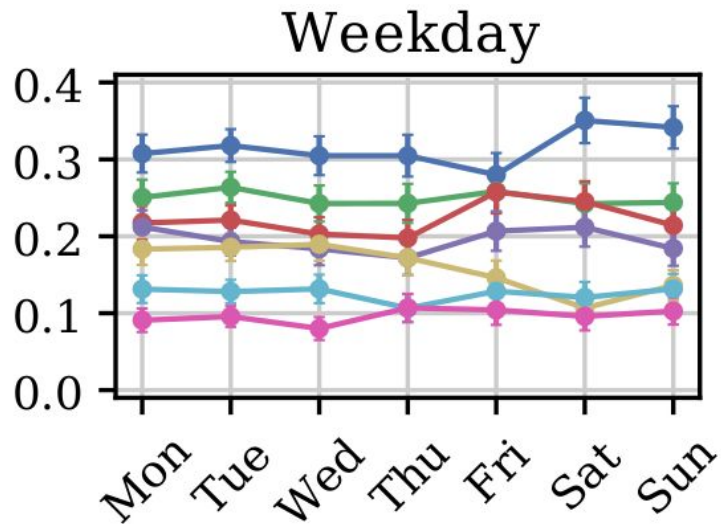
Prior knowledge



Motivation



Motivation day and time



- media
- intrinsic learning
- conversation
- bored/random
- work/school
- current event
- personal decision

Correlation: Motivation *vs.* information need

information need motivation	fact	in-depth	overview	sig.
media	0.38 (1.00)	0.19 (0.87)	0.43 (1.12)	***
intrinsic learning	0.29 (0.76)	0.35 (1.62)	0.35 (0.92)	***
conversation	0.43 (1.13)	0.20 (0.93)	0.36 (0.94)	***
bored/random	0.31 (0.83)	0.23 (1.05)	0.45 (1.17)	***
work/school	0.39 (1.04)	0.23 (1.09)	0.36 (0.93)	
current event	0.36 (0.95)	0.28 (1.30)	0.35 (0.92)	***
personal decision	0.32 (0.85)	0.29 (1.35)	0.38 (0.97)	***
response perc.	0.38	0.21	0.39	

Correlations: Motivations *vs.* prior knowledge

prior knowledge motivation	familiar	unfamiliar	sig.
media	0.42 (0.83)	0.58 (1.22)	***
intrinsic learning	0.57 (1.14)	0.41 (0.87)	***
conversation	0.49 (0.98)	0.49 (1.04)	***
bored/random	0.53 (1.07)	0.45 (0.95)	
work/school	0.52 (1.04)	0.46 (0.97)	
current event	0.52 (1.03)	0.46 (0.98)	
personal decision	0.50 (0.99)	0.48 (1.02)	
response perc.	0.50	0.47	

Conclusions

- English Wikipedia is consulted for a variety of use cases and none are dominant.
- Shallow information needs (overview + lookup = 77%) appear to be more common than deep information needs (21%).
- Readers have nearly identical shares in being familiar (50%) vs. unfamiliar (47%) with the topic of interest.
- Extrinsic vs. intrinsic:
 - Extrinsic triggers: media (30%), conversation (22%), work/school (16%), current event (13%).
 - Intrinsic triggers: intrinsic learning (25%), bored (20%), personal decision (10%)

Behavioral patterns associated with
different use case

Features

Survey

Motivation

Information need

Prior knowledge

Request

Country

Continent

Local time weekday

Local time hour

Host

Referer class

Article

In-degree

Out-degree

Pagerank

Text length

Pageviews

Topics

Topic entropy

Session/Activity

Session length

Session duration

Average dwell time

Average pagerank difference

Average topic distance

Referer class frequency

Session position

Number of sessions

Number of requests

Subgroup discovery

- Each survey question-response forms a *target*. Consider work/school motivation, for example.
- For each target, we do rule mining to detect behavioral patterns that are significantly different than the rest of the population, e.g., a larger share of long sessions compared to the whole population.

Subgroup analysis - Information need

- More homogenous subgroups with some notable exceptions.
- Users from Asia describe their information needs significantly more often as in-depth (more investigation needed)
- Obtaining overview is more common among desktop users
- Fact checking is more often observed in *sports*

Subgroup analysis - Prior knowledge

Users are familiar with

- topics that are more spare-time oriented (*sports, 21st century, TV, movie, and novels*)
- topics that are popular (many pageviews)
- articles that are longer, and are more central in the link network.

Subgroup analysis - Motivation

(a) T : “motivation = work/school”; $P(T) = 19.5\%$

Subgroup S	$P(S)$	$P(S T)$	$P(T S)$	lift	sig.
topic (mathematics)	7.9%	17.1%	34.8%	2.17	***
topic (war, history)	4.4%	9.6%	34.7%	2.16	***
topic (technology)	13.2%	23.7%	28.8%	1.79	***
topic (biology, chemistry)	8.6%	14.0%	26.2%	1.63	***
host = desktop	35.5%	57.8%	26.1%	1.63	***
article pagerank ≥ 9.98	20.0%	32.4%	26.1%	1.62	***
avg. time difference ≥ 9.40	7.7%	11.5%	24.0%	1.50	***
avg. pagerank difference < -4.35	7.6%	11.2%	23.6%	1.47	***
topic (literature, art)	10.1%	14.7%	23.5%	1.46	***
avg. time difference: [3.60:9.40[7.7%	11.0%	23.1%	1.44	***
num. (referrer=search) ≥ 2	20.5%	28.5%	22.4%	1.39	***
session duration > 6.60	18.0%	24.2%	21.6%	1.34	***

(b) T : “motivation = bored/random”; $P(T) = 16.1\%$

Subgroup S	$P(S)$	$P(S T)$	$P(T S)$	lift	sig.
referrer class: internal	9.4%	14.0%	29.0%	1.49	***
num. of requests ≥ 8	11.8%	16.6%	27.5%	1.41	***
topic (sports)	5.9%	8.0%	26.1%	1.34	**
num. (referrer=internal) ≥ 1	17.1%	22.7%	25.9%	1.33	***
session position: [0.33:0.75[7.5%	9.8%	25.6%	1.31	**
avg. topic distance (session) ≥ 1.08	7.5%	9.8%	25.2%	1.29	*
topic (21st century)	25.1%	32.1%	25.0%	1.28	***
session length ≥ 3	22.2%	28.3%	24.8%	1.27	***
avg. time difference: [0.68:1.56[7.7%	9.7%	24.7%	1.27	*
num. (referrer=none) ≥ 2	9.7%	12.2%	24.5%	1.26	*
topic (tv, movies, novels)	34.1%	41.4%	23.7%	1.21	***
# article pageviews ≥ 63606	19.8%	23.5%	23.1%	1.19	**

Let's step back and summarize

- Built a taxonomy of Wikipedia readers (information need, prior knowledge, and motivation).
- Quantified the prevalence and interaction between the use cases.
- Studied the survey responses in the context of different sessions, articles, and requests.

Implications and future directions

- Robustness of results across languages
- Predicting motivation at the reader level
- Predicting motivation at the article level
- What is the Movement's role?
 - 21% in-depth reading
 - Motivation variations and learning
 - Screen size and content
 - ...
- What else could we do?

References

- S. Goel, J. M. Hofman, and M. I. Sirer. Who does what on the Web: A large-scale study of browsing behavior. In International Conference on Web and Social Media, 2012.
- R. Kumar and A. Tomkins. A characterization of online browsing behavior. In International Conference on World Wide Web, 2010.
- A. Broder. A taxonomy of web search. In ACM SIGIR Forum, 2002.
- D. E. Rose and D. Levinson. Understanding user goals in web search. In International Conference on World Wide Web, 2004.
- I. Weber and A. Jaimes. Who uses web search for what: and how. In International Conference on Web Search and Data Mining, 2011.
- A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In Workshop on Web Mining and Social Network Analysis, 2007.
- H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In International Conference on World Wide Web, 2010.
- T. Ryan and S. Xenos. Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage. *Computers in Human Behavior*, 27(5):1658–1664, 2011.
- O. Arazy, H. Lifshitz-Assaf, O. Nov, J. Daxenberger, M. Balestra, and C. Cheshire. On the “how” and “why” of emergent role behaviors in Wikipedia. In Conference on Computer-Supported Cooperative Work and Social Computing, 2017.

References (cont'd)

- O. Nov. What motivates Wikipedians? *Communications of the ACM*, 50(11):60–64, 2007.
- D. Jurgens and T.-C. Lu. Temporal motifs reveal the dynamics of editor interactions in Wikipedia. In *International Conference on Web and Social Media*, 2012.
- J. Lehmann, C. Müller-Birn, D. Laniado, M. Lalmas, and A. Kaltenbrunner. Reader preferences and behavior on Wikipedia. In *Conference on Hypertext and Social Media*, 2014.
- J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Characterizing and modeling the dynamics of online popularity. *Physical Review Letters*, 105(15):158701, 2010.
- A. Spoerri. What is popular on Wikipedia and why? *First Monday*, 12(4), 2007.
- V. Waller. The search queries that took Australian Internet users to Wikipedia. *Information Research*, 16(2), 2011.
- D. Lamprecht, D. Dimitrov, D. Helic, and M. Strohmaier. Evaluating and improving navigability of Wikipedia: A comparative study of eight language editions. In *International Symposium on Open Collaboration*, 2016.
- A. Paranjape, R. West, L. Zia, and J. Leskovec. Improving website hyperlink structure using server logs. In *International Conference on Web Search and Data Mining*, 2016.
- P. Singer, D. Helic, B. Taraghi, and M. Strohmaier. Detecting memory and structure in human navigation patterns using Markov chain models of varying order. *PloS One*, 9(7):e102070, 2014.
- R. West and J. Leskovec. Human wayfinding in information networks. In *International Conference on World Wide Web*, 2012.

Thank you! :)

Credits

- Page 1: Note that the logos used on the first slide belong to the corresponding institutions.
- Page 10: Leonard Cohen, 1988 01 by Gorupdebesanez,
https://commons.wikimedia.org/wiki/File:Leonard_Cohen,_1988_01.jpg
CC BY-SA 3.0 <https://creativecommons.org/licenses/by-sa/3.0/deed.en>