Leveraging Behavioral Data to Improve Workplace Productivity

Ahmed Hassan Awadallah
http://aka.ms/ahmed
What makes Information Personal?

Owned by “me”

About “me”

Directed toward “me”

Sent by “me”

Relevant to “me”

Experienced by “me”

Jones, 2008, Keeping Found Things Found
Enormous Volume of Information Exchanged in the Workplace

In 2018, the total number of business emails sent and received per day is 120.4 billion by over 1 billion users.

Average office worker receives 76 non-spam emails a day, sends out 32 business emails a day.

Source: Email Statistics Report. The Radicati Group INC.
Significant time spent interacting with this information

Source: International Data Corporation (IDC), McKinsey Global Institute analysis
People should spend less time searching for and organizing information and more time acting on and drawing insights from it.
Email requesting a document

**Updated Proposal**

“can you share the latest version of the proposal?”

---

Email requesting a meeting

**Re: Updated Proposal**

“The proposal looks great. Let’s meet to discuss the next steps next week?”

---

Meeting Notes

Carlos Slattery

Tue 7/28/2018 5:29 PM

You 😊

Saw your notes. Would you mind sharing your notebook with the group before our next discussion?

Thanks,

Carlos

Sure thing. I'm on it. I'll get back to you.

---

Recommended

[Images of Microsoft Teams](https://www.microsoft.com)

[Images of Microsoft Teams](https://www.microsoft.com)
The World Wide Web

Trillions of pages indexed.
Billions of queries per day.
Modem Web Search Engines

User Behavioral Data
What is Behavioral data?

Behavioral logs are *traces of human behavior*

- seen through the lenses of whatever sensors we have
- Examples: utterance, queries, browsing, invoking apps, clicks on UI elements, keystrokes, gaze patterns, physiological responses
How can behavioral data help?

- Understand users’ intents and needs
- Learn to build better ML models
- Measure user satisfaction and engagement
How is Behavioral data used in Web Search?

Understand Tasks and Intent from Query Logs

- Online check credit
- Mortgage in principle
- Quit smoking benefits
- Solicitors near me

- Loans for house
- Facebook
- House buying guide
- Houses for sale

Session 1:
- 10:00am
- 10:03am
- 10:07am

Session 2:
- 12:30pm

Session 3:
- 17:00pm
- 17:02pm
- 17:06pm

Session 4:
- 18:15pm

[Jones and Klinkner, 2008]
How is Behavioral data used in Web Search?

Learn to Rank using the Click Graph

[Joachims, 2002; Craswell and Szummer, 2007]
How is Behavioral data used in Web Search?

Measure User Satisfaction

[Satisfied Trails]
Trail 1: Q 4s RL 1s SR 53s SR 118s END
Trail 2: Q 3s Q 3s SR 10s AD 44s END
Trail 3: Q 4s RL 1s SR 53s SR 118s END
Trail 4: Q 3s Q 5s SR 10s AD 44s END
Trail n: Q 4s RL 1s SR 53s SR 118s END
Trail n-1: Q 3s Q 5s SR 10s AD 44s END

[ Dissatisfied Trails ]
Trail 1: Q 4s RL 1s SR 53s SR 118s END
Trail 2: Q 3s Q 3s SR 10s AD 44s END
Trail 3: Q 4s RL 1s SR 53s SR 118s END
Trail 4: Q 3s Q 5s SR 10s AD 44s END
Trail n: Q 4s RL 1s SR 53s SR 118s END
Trail n-1: Q 3s Q 5s SR 10s AD 44s END

[Hassan et al., 2010]
Improving Productivity by leveraging Behavioral Data
Three Productivity Applications

**Communications**
Help users respond to and take actions on email messages

**Documents**
Surface document recommendation relevant to the user’s context

**Tasks**
Assist users with task management and planning
Email Intelligence
The Lifetime of an Email Message

Different lifetime spans and different interaction patterns

Revisiting to **take an action** or **retrieve information**
Many messages are deleted upon receipt, others are only read.

Responding to a message is a frequent event and is most likely to happen during the first few visits.
Let’s focus on responding to a message

If the message is requesting an action from the recipient, it is more likely to receive a reply.

But it is also more likely to receive a slow reply.
Message to Action Recommendation

Email requesting a document

Updated Proposal

“can you share the latest version of the proposal?”

Share Document  Open Email

Email requesting a meeting

Re: Updated Proposal

“The proposal looks great. Let’s meet to discuss the next steps next week?”

Create Meeting  Open Email
Why do we send email messages?
Why do we send email messages?

Information Seeking/Sharing
Why do we send email messages?

Scheduling/Coordinating
Why do we send email messages?
Message to Action: Understanding Intent
Message to Action: Understanding Intent
Message to Action: Recommendation

Wang et al., SIGIR 2019
People frequently exchange questions/answers over email and they often need to go back to such information — more than 5 times a month for 80% of them

How often do you need to go back to threads because they contain Questions/Answers?

Yang et al., SIGIR 2018
Many of them receive the same question and have to answer it over and over again
— more than 5 times a month for 33% of them.
Question/Answer Exchange over Email

How-to and When/Where/Who questions are the most popular types people need to get back to

Yang et al., SIGIR 2018
Hi Alice,

I am trying to plan for the next round of ABC testing. Do either of you know what percentage of the ABC tests are automated and how long does it take to go through an entire cycle of tests if there are no bugs?

Thank you!

Thanks,  
Mike

Similar questions you answered before:

Do you know what percent of the ABC tests are manual and how much time does it take to run them?

It takes two day to run through all the tests, barring any problems. Most of the old tests are fully automated but we have been adding new ones that are still manual. So approximately 20% of the testing is still manual.
Question/Answer Exchange over Email

Learn to identify question/answer pairs using content and user interaction information

Yang et al., AAAI 2019
Document Recommendation
Document Recommendations

Office 365

Apps
- Outlook
- OneDrive
- Word
- Excel
- PowerPoint
- OneNote
- SharePoint
- Teams
- Yammer

Documents

Recommended
- You edited this Thu at 1:13 PM
- Jack share this file with you Wed at 1:41 AM
- Team commented on this file Mon at 3:25 PM
- You recently opened this Jun 20

Data for the WWW 2020 Paper microsoft/my.sharepoint...
Slides for the WWW 2020 Paper microsoft/my.sharepoint...
Draft of the WWW 2020 Paper microsoft/my.sharepoint...
Sample of a WWW 2019 Paper microsoft/my.sharepoint.com

Recent
- Slides for the research group meeting OneDrive for Business
- Slides for brain storming OneDrive for Business

Last opened by you: 7h ago
Sharing: Shared
Activity: Amy edited 6h ago

Recommended Document Pane (RDP)
### Types of Behavioral Data

<table>
<thead>
<tr>
<th></th>
<th>Observational</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lab Studies</strong></td>
<td>In-lab behavior observations</td>
<td>In-lab controlled tasks, comparisons of systems</td>
</tr>
<tr>
<td>Controlled tasks</td>
<td><strong>Observational</strong></td>
<td><strong>Experimental</strong></td>
</tr>
<tr>
<td>in laboratory,</td>
<td>In-lab behavior observations</td>
<td>In-lab controlled tasks, comparisons of systems</td>
</tr>
<tr>
<td>with detailed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>instrumentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Panel Studies</strong></td>
<td>Ethnography, case studies, panels (e.g., Nielsen)</td>
<td>Clinical trials and field tests</td>
</tr>
<tr>
<td>In the wild,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>real-world tasks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ability to probe</td>
<td></td>
<td></td>
</tr>
<tr>
<td>for detail</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Log Studies</strong></td>
<td>Logs from a single system</td>
<td>A/B testing of alternative systems or algorithms</td>
</tr>
<tr>
<td>In the wild,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no explicit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>feedback but lots</td>
<td></td>
<td></td>
</tr>
<tr>
<td>of implicit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>feedback</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Types of Behavioral Data

<table>
<thead>
<tr>
<th>Lab Studies</th>
<th>Observational</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controlled tasks, in</td>
<td>In-lab behavior observations</td>
<td>In-lab controlled tasks, comparisons of systems</td>
</tr>
<tr>
<td>laboratory, with detailed instrumentation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel Studies</th>
<th>Ethnography, case studies, panels (e.g., Nielsen)</th>
<th>Clinical trials and field tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the wild, real-world tasks, ability to probe for detail</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Log Studies</th>
<th>Logs from a single system</th>
<th>A/B testing of alternative systems or algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the wild, no explicit feedback but lots of implicit feedback</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Types of Behavioral Data

<table>
<thead>
<tr>
<th>Lab Studies</th>
<th>Observational</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controlled tasks, in laboratory, with detailed instrumentation</td>
<td>In-lab behavior observations</td>
<td>In-lab controlled tasks, comparisons of systems</td>
</tr>
<tr>
<td>Panel Studies</td>
<td>Ethnography, case studies, panels (e.g., Nielsen)</td>
<td>Clinical trials and field tests</td>
</tr>
<tr>
<td>In the wild, real-world tasks, ability to probe for detail</td>
<td>Logs from a single system</td>
<td>A/B testing of alternative systems or algorithms</td>
</tr>
<tr>
<td>Log Studies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In the wild, no explicit feedback but lots of implicit feedback</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Lab Study

Contextual inquiry, where the participants were asked to look through their document recommendations and share some information about it.

Variety of questions with free form text answers to understand *what* and *why* they interact with recommendations.
Field Study

A Brief History of the Middle Earth
Author is Samwise Gamgee
SharePoint

Shared with you by Gandalf the Grey

Overview of past actions
- Last time, you edited this
  2 days ago
- Charlie Davidson and you edited this
  4 hours ago
- Eve Frank, Charlie Davidson and 4 others commented on this
  3 days ago
- Eve Frank, Ivan Judy and 2 others mentioned you
  a day ago

Do you recognize this document?
- NO
- YES

Had you planned on accessing this file today or in the near future?
- Yes
- No

Now that you have seen the document, would you be interested in accessing it?
- Yes
- No
Field Study

100+ PARTICIPANTS

~2000 DOCUMENTS

- Recency of Access
- Richness of Interactions
- Presentation

Do you recognize this document?

Yes

Had you planned on accessing this file today or in the near future?

No

No

Now that you have seen the document, would you be interested in accessing it?
Familiar documents are easier and faster to recognize compared to less familiar ones.
Two distinct intents: **refinding** known and recently accessed document and **discovering** new documents.
Richer presentations are helpful (verbatim feedback) but take more time to process.
Large-Scale Log Study

Visit 1: Just visit
Visit 2: Open a file
Visit 3: Direct to an app
Visit 4: Open multiple files
Visit 5: Open files and direct to an app

Visit office.com
Open A File
Redirect To Other Apps

800K USERS
8.3M VISITS

Xu et al., WWW 2020
Document Recommendations: How to measure utility?

<table>
<thead>
<tr>
<th></th>
<th>Shown in RDP</th>
<th>Not shown in RDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open in RDP</td>
<td>47.5%</td>
<td>—</td>
</tr>
<tr>
<td>Open Elsewhere</td>
<td>17.2%</td>
<td>35.3%</td>
</tr>
</tbody>
</table>

Click Through Rate (CTR) = \( \frac{\text{Number of Opening a File from RDP}}{\text{Number of Visits}} \)

Time to Open (TTO) = Time Elapse between Visit and File Open from RDP

Recognize Rate (RR) = \( \frac{\text{Number of Opening a File from RDP}}{\text{Number of Opening a File shown in RDP}} \)
Document Recommendations: Interactions

- CTR decrease from the Position 1 (leftmost) to 4 (right most)
- Most of the people do not visit the “next page” of the RDP, but once they go there, there is a jump up of the CTR
Large-Scale Log-Based Experimental Study: Document Explanations

(a) Pairwise Comparison on CTR

(b) Pairwise Comparison on RR
Task Intelligence
Challenges in Task Management

- Intelligent systems (digital assistants, etc.) store / remind users about tasks

- Tasks can be explicitly specified or inferred

- Users still face several challenges. For example:
  1. Task Scheduling – Duration Estimation
  2. Task status Tracking – Task Auto Deprecation
User Interactions as labels for ML Models

**User Interactions**

- $(x_{u}, y_{u})$
  - **Personal History**

- $(x_{g}, y_{g})$
  - **Collective Behavior**

Train **Task-specific model**
Task Duration Estimation

• Time estimation occurs often (every task needs this)

• People struggle to estimate how long tasks take, esp. for new tasks

• Forecast durations using large-scale data

Aspirational goal is building such experience
Duration Estimation is hard even for people

- Duration increases with recurrence
- Connections to planning fallacy, etc.

Changes in mean task duration with increasing task occurrence ($i$) ($\pm$SEM)

$R^2 = 0.7052$
Learning from Personal & Collective User Behavior Data

- Anonymized appointment data
- >3M appointments from 700k users
- Appt subjects are hashed
- Time durations blocked by users
- Grouped into four classes
Personal behavior data is important when available

- History is important

- Locations correlate with longer durations while phone numbers correlate with shorter ones

- More basic language is used to describe clearer/simpler tasks

Generated task features:

<table>
<thead>
<tr>
<th>Attribute</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean duration (per user-task)</td>
<td>+0.41487</td>
</tr>
<tr>
<td>Median duration (per user-task)</td>
<td>+0.41056</td>
</tr>
<tr>
<td>Mean duration (per task)</td>
<td>+0.35767</td>
</tr>
<tr>
<td>Median duration (per task)</td>
<td>+0.35035</td>
</tr>
<tr>
<td>Mean duration (per user)</td>
<td>+0.32666</td>
</tr>
<tr>
<td>Median duration (per user)</td>
<td>+0.30167</td>
</tr>
<tr>
<td>Has location</td>
<td>+0.23894</td>
</tr>
<tr>
<td>Std deviation duration (per user-task)</td>
<td>+0.14959</td>
</tr>
<tr>
<td>Std deviation duration (per user)</td>
<td>+0.14557</td>
</tr>
<tr>
<td>Maximum token length</td>
<td>+0.05715</td>
</tr>
<tr>
<td>Has country</td>
<td>+0.05117</td>
</tr>
<tr>
<td>Has address</td>
<td>+0.05024</td>
</tr>
<tr>
<td>Minimum token length</td>
<td>-0.05258</td>
</tr>
<tr>
<td>Start minute</td>
<td>-0.05719</td>
</tr>
<tr>
<td>Has phone number</td>
<td>-0.06061</td>
</tr>
<tr>
<td>Fraction text stop words (a, the, etc.)</td>
<td>-0.07673</td>
</tr>
<tr>
<td>Total number of stop words in text</td>
<td>-0.08378</td>
</tr>
<tr>
<td>Number of unique stop words in text</td>
<td>-0.08741</td>
</tr>
<tr>
<td>Task popularity (across all users)</td>
<td>-0.13104</td>
</tr>
<tr>
<td>Number of action verbs</td>
<td>-0.14705</td>
</tr>
</tbody>
</table>

† Non-(time/duration/history) attributes are based on task description

$r = \text{correl with appt duration}$
Learning to efficiently predict task duration

- Performance changes with:
  - + more data
  - + more features

---

**Graph:**

- **Micro Accuracy**
- **X-axis:** 25K, 50K, 100K, 200K, 400K, 800K, 1.6M, Full
- **Y-axis:** 50%, 55%, 60%, 65%, 70%, 75%

- **Lines:**
  - DL-Content (red)
  - DL-TextOnly (orange)
  - LR-Content (blue)
  - LR-TextOnly (light blue)

---

**Bar Chart:**

- **Micro Accuracy**
- **X-axis:** Context, History, Content, Content+Context, Content+History, Content+History+Context
- **Y-axis:** 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%

- **Values:**
  - Content: 50.30%
  - History: 62.39%
  - Content + Context: 72.23%
  - Content + History: 72.45%
  - Content + History + Context: 78.92%
Task Auto-Deprecation

- Show pending tasks
- Flag or deprecate completion candidates
- Suppress notifications
- Other applications possible, incl. task ranking, task prioritization, etc.

Pending Tasks

“I’ll work on that later.”
Sent to: Gregg Newton — 8/21/2018, 12:43pm

Snooze | View email | Completed

“I will find out what else they have.”
Sent to: Clayton Jones — 8/25/2018, 09:01am

Snooze | View email | Completed

It looks like this task is already complete ...

“I will send you the file by end of day.”
Sent to: Norma Saunders — 8/16/2018, 10:54am

Snooze | View email | Completed | Not completed
Learning from User Behavior Data

- Offer a feedback affordance for users to indicate task completion
  - “Complete” clicks help form ground truth
    - Only says task was completed **BY some time**, not **WHEN** the task completion occurred
Task Completion Over Time

- Compute fraction of tasks completed at $t_n$, all tasks and per task type
  - Task type by priority (high-pri language) and by activity (call, email, investigate)

High priority tasks are completed faster

Relative completion timing: Call < Email < Investigate

Connected to avg relative complexity

Some email tasks can be handled as quickly as a phone call
Learning to efficiently predict task completion

• Five feature classes:
  – **Time**: time elapsed since task created
  – **Content**: n-grams, verbs, priority, due date, is conditional, intent, etc.
  – **Email**: subject n-grams (no email body), is reply, number of recipients, etc.
  – **Notifications**: logged Cortana notifications (16% of tasks), num notifications, etc.
  – **User History**: (38% of users), historic tasks, completion time/rates, etc.

Leverage both content and behavioral features
Learning to efficiently predict task completion

- Strong performance and improves with more data
Challenges and Opportunities
You get what you measure
Feedback Loop

Experience

Data

Engagement
Privacy powers Trust
Trust enables Data
Data powers Innovation
Contributors: Ahmed Awadallah, Paul Bennett, Susan Dumais, Saghar Hosseini, Subho Mukhrejee, Farheen Omar, Bogdan Popp, Robert Rounthwait, Robert Sim, Milad Shokouhi, Wei Wang, Ryen White, Guoqing Zheng

Interns: Farnaz Jahanbakhsh, Yang Liu, Yang Xiao, Orson Xu
Thank you

http://aka.ms/ahmed