Panmodal Information Interaction

RYEN W. WHITE, Microsoft Research, Redmond, WA, USA
CHIRAG SHAH, University of Washington, Seattle, WA, USA

The emergence of generative artificial intelligence (GenAI) is transforming information interaction. For decades, search engines such as Google and Bing have been the primary means of locating relevant information for the general population. They have provided search results in the same standard format (the so-called "10 blue links"). The simplicity and consistency of this user experience has been satisfying and predictable for many search engine users. Those users have consumed results and derived their own answers from those results. The recent ability to chat via natural language with AI-based agents and have GenAI automatically synthesize answers in real-time (grounded in top-ranked results) is changing how people interact with and consume information at massive scale. These two *information interaction modalities* (traditional search and AI-powered chat) coexist in current search engines, either loosely coupled (e.g., as separate options/tabs) or tightly coupled (e.g., integrated as a chat answer embedded directly within a traditional search result page). We believe that the existence of these two different modalities, and potentially many others, is creating an opportunity to re-imagine the search experience, capitalize on the strengths of many modalities, and develop systems and strategies to support seamless flow between them. We refer to these as *panmodal* experiences.[1] Unlike monomodal experiences, where only one modality is available and/or used for the task at hand, panmodal experiences make multiple modalities available to users (multimodal), directly support transitions between modalities (crossmodal), and seamlessly combine modalities to tailor task assistance (transmodal). While our focus is search and chat, we also present a more general vision for the future of information interaction using multiple modalities enabled by the emergent capabilities of GenAI and grounded in several already-available technologies.

ACM Reference Format:
Ryen W. White and Chirag Shah. 2024. Panmodal Information Interaction. 1, 1 (November 2024), 7 pages.

1 INTERACTING WITH INFORMATION

Information is essential for effective decision making and action. Information systems such as search engines facilitate rapid and comprehensive access to that information. An information interaction modality describes the particular mode of how people engage with such a system, including different interaction paradigms (e.g., query-response, multi-turn dialog, proactive suggestions), our focus here, but also different input mechanisms (e.g., text, speech, touch), and different device types. For decades, mainstream search interfaces in search engines have offered just one information interaction modality: query-response or simply *search* (comprising text query, hyperlinked search results, click-through to landing pages, and iterative query refinement as needed to find relevant information). While search engines have been a primary source of information for consumers, emerging modalities fueled by frontier models such as Google's

[1] An alternative nomenclature would be *multimodal* experiences. Multimodal is a common term in human-computer interaction for the availability of multiple modalities, but is also mostly used in AI these days to refer to the use of different content types (text, images, video, audio, and so on) in foundation model training and inference. As we will discuss, we also believe that multimodal is a subset of panmodal.

Authors' addresses: Ryen W. White, ryenw@microsoft.com, Microsoft Research, Redmond, WA, USA; Chirag Shah, chirags@uw.edu, University of Washington, Seattle, WA, USA.

Gemini and OpenAI's GPT-4o, that can reason over text, audio, and vision inputs and generate text and multimedia outputs, can complement search engine capabilities and have created new possibilities for information interaction [11].

2 INTERACTION MODALITIES

The *chat* interface is an essential component of many GenAI-based systems. Multi-turn dialog has long shown promise as a way to engage with information systems [5], but is now going mainstream in support of complex tasks via progress in GenAI [15] and in GenAI-based conversational systems such as ChatGPT[2] and Pi.[3] SearchGPT, recently trialed by OpenAI, provides highly relevant, verifiable answers in a conversational experience. Search engines can now show GenAI answers directly on result pages—minimizing user effort in examining search results but also removing human control over answer generation [9], which can have its own drawbacks (e.g., fewer learning opportunities)—and let users follow up via multi-turn conversation for clarification or to seek additional information.

Traditional search still has utility for some tasks, e.g., for fact finding or navigational tasks, and may be preferred by some searchers given its focus on providing information sources directly rather than synthesized answers. GenAI is also prone to hallucinate (i.e., generate nonsensical or inaccurate outputs), making sole reliance on its generated answers inadvisable, although source attribution and answer verification are now creeping in to help users better assess what they can use and trust.

When only one modality is available and/or used (e.g., search only, chat only), we call it *monomodal information interaction*. When multiple modalities are used separately or collectively for accessing information, we refer to it as *panmodal information interaction*. Information systems with both search and chat functionality may only be the beginning. With the advent of GenAI models that can, among other things, disambiguate user intents, create new interfaces, and surface various types of information, we foresee considerable growth opportunity in panmodal experiences.

3 PANMODAL INFORMATION INTERACTIONS

Panmodal experiences offer multiple modalities but also provide assistance in moving between them or combining them for more effective task completion (Figure 1). In monomodal interactions, people select just one modality for the task (e.g., the only modality available or feasible to use in the context, or selected from alternatives). Search may be preferred for tasks with broad coverage online (e.g., personal tasks such as finding recipes, planning a vacation, and shopping), whereas chat may be preferred for more nuanced tasks that require specialized knowledge, iteration, and the creation of new content (e.g., professional tasks such as background research, report writing, and coding).

3.1 Multimodal Information Interactions

Multimodal interfaces have been studied for decades in the human factors community [6], including for information interactions. Search systems have long provided different facilities and search strategies based on user needs and goals [2] and even suggested specific search engines given the current query [13]. As mentioned earlier, search engines now offer multiple modalities, search and chat, juxtaposed in the search interface. This is available on Bing and in Google as the so-called search generative experience (SGE), recently promoted by Google as *AI Overviews*.[4] Bing now offers both a Generative Search experience,[5] with embedded GenAI answers on the search engine result page, and

[2] https://chat.openai.com/
[3] https://pi.ai
[4] https://blog.google/products/search/generative-ai-google-search-may-2024/
[5] https://blogs.bing.com/search/july-2024/generativesearch/

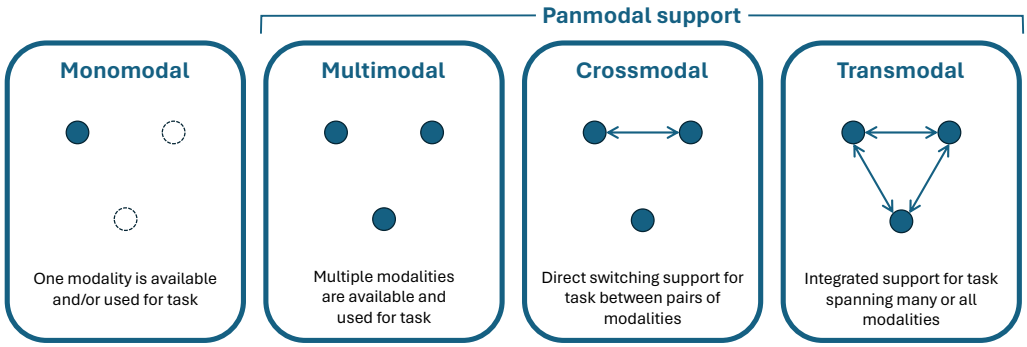

Fig. 1. Examples of different modality usage patterns and panmodal support for information interaction.

Deep Search,[6] utilizing GPT-4 to create a more complete representation of users' search intent, potentially comprising multiple internal search queries, and an associated ranking of their merged search results.

For users, making informed selections about which modality to use also depends on sound mental models of system capabilities. People have developed mental models of search engines over decades of use. For chat-based systems, users' mental models are still being formed and users may experiment with different prompt formulations to tailor the system output to meet their expectations. Showing both GenAI-powered chat experiences and traditional search results in a single result page, as in SGE, is also imprecise and can be overwhelming for users, potentially impeding task progress. The use of expensive foundation models is also unnecessary if the results alone are sufficient to answer the question.

3.2 Crossmodal Information Interactions

While users have preferences for a modality given a task or their prior experiences, there are times when their initial choice does not pan out, or they may simply require some of the capabilities of a different modality, and they need to switch. We refer to these as *crossmodal information interactions*. Crossmodal has been explored in the context of search (e.g., text and voice-based input/output, different device types [4]), but transitions between search and chat are less well understood and supported by current information systems. Switches may happen for a variety of reasons (e.g., search to chat: search is insufficiently interactive and not tailored to user needs, chat to search: users need more comprehensiveness and more diverse perspectives).

We note that current mainstream search experiences are multimodal with very limited crossmodal support (e.g., GenAI-powered chat on result pages). An information system could engage with users to help with this switching decision and suggest the most suitable modality for the current task or even task step/stage. Switching considerations also extend beyond the purview of technological solutions, e.g., involve modality awareness, which can alter usage habits, and connected to time spent with the chat modality, in order to learn how best to use its functionality.

3.3 Transmodal Information Interactions

As has been clear thus far, there is no one modality of information interactions that could satisfy users for the full range of their informational needs. More importantly, using multiple modalities strategically could help users not only

[6] https://blogs.bing.com/search-quality-insights/december-2023/Introducing-Deep-Search/

accomplish their tasks more effectively, but also unlocks possibilities for tasks that are not traditionally performed through monomodal, multimodal, or even crossmodal approaches. These are *transmodal information interactions*.

Users could be doing transmodal interactions for complex information needs that require more dynamic support (e.g., for using several different modalities for the same task, either sequentially [4] or concurrently [12]; representing and preserving task state; using GenAI to decompose tasks, plan actions, select and sequence modalities, and so on). Examples of domains and tasks for which we might expect to see or desire transmodal interactions include healthcare (e.g., diagnosis and mitigation) or research (e.g., creating a diverse set of summary reports for a business need).

However, given that most current interfaces lack a clear support for running concurrent modalities with seamless integration and focus on the task at hand, current users of search systems could be doing such transmodal work through ad hoc workflows. Considering the importance of supporting transmodal information interactions and current lack of such support, there is a need for platforms to not only facilitate the use of multiple modalities, but also to create user experiences with the most appropriate constellation of modalities and connections between them.

4 PANMODAL FUTURES

Going forward, we expect that panmodality will play more of a central role in information interaction. There are clear benefits to developing specialized modalities with specific capabilities, in their complimentarity, and in efforts to bring them together as needed for tasks. Future information systems may well be built around different modalities in a unified, seamless way, with accommodations for adapting system operation and the user experience when additional personal/contextual information is available, e.g., require shorter or even no manual requests while offering proactive experiences [3] when the AI is confident about the task or intent. Different modalities can be offered within a single application or by using a federated approach, surfacing modalities provided by a range of different applications, including those developed by third parties. Tools such as Google Dialogflow, Amazon Lex, and Microsoft Copilot Studio can be used to create conversational interfaces that seamlessly switch between different modalities such as text, voice, and chat, depending on the user's context and preferences.

Panmodal experiences will utilize different information interaction modalities depending on user preferences and/or the nature of the search task; users can focus on tasks and the system selects or suggests the most appropriate modality or modality sequence. This requires task modeling and new routing mechanisms capable of matching tasks to relevant modalities. Improvements to personalization, intent understanding, and context awareness will yield more accurate and appropriate modality selection. For example, a user could start a conversation via the audio interface in an automobile, continue that with touch interface on their smartphone, and then explore the options through gestures on a wall display – all driven by human initiative and/or guided by (semi-)autonomous GenAI agents such as Taskweaver [7]. The modalities could be served by different providers and the selections and transitions could be brokered by agents engaging with users and/or other agents [14], e.g., to provide diverse perspectives on the available options.

The primary focus of the interaction will be on users and their needs/tasks, while the agent figures out the best way to move forward with and accomplish their tasks in different environments and with appropriate task assistance. As part of this, systems will need to support switching between modalities while preserving context.

There are additional information interaction modalities beyond search and chat. For example, bespoke interfaces generated natively by GenAI for the task at hand (e.g., such support for personalized and task-specific interface generation is already provided by Google Gemini and other multimodal models), interactive data visualizations created by GenAI, such as those in Tableau Pulse, akin to dynamic queries in the research literature [1], proactive recommendations from GenAI based on audio and vision sensing (e.g., unlocked using multimodal capabilities such as those in Amazon's

Rekognition or enabled end-to-end via Google's Project Astra), and moving from information access to information use, in GenAI-based operation or suggestion of tools to support task completion (e.g., via action models and agents such as Adept (recently acquired by Amazon) and MultiOn) [8] (with human involvement and oversight). This will also expand beyond interaction paradigms into new modes of interaction (e.g., tactition, gesture, eye gaze) and new device types (e.g., smart rings such as the Oura and Samsung Galaxy Ring, smart glasses such as the joint offering from Ray-Ban and Meta), where GenAI could help interpret signals and add intelligence. This may also include reducing the interaction cost per the reliability of the contextual information available at the time. Determining the most appropriate type of panmodality support could serve as useful reminders to users even if we do not directly support their use in-situ in the application or support transitions to them in other applications. For example, for a task where transmodal support is best suited, GenAI could enumerate potential modalities and candidate sequence orders per a generated plan, an activity which most, if not all, state-of-the-art foundation models now find to be fairly straightforward. This could also be used as a way to develop user awareness of different modalities available to them and help evolve their usage habits over time to make better use of those options on their own.

Beyond more efficient and effective information interaction and better task outcomes, there are other implications for information systems from utilizing multiple modalities, e.g., cost (lower costs if non-GenAI solutions can be employed for some tasks), evaluation (need richer interaction models for panmodal experiences), trust (to perform actions, generate answers, etc.). There are implications for the economics of AI from managing costs and revenue generation across modalities as well as selecting and ranking modalities across many first- and third-party providers. Each modality may have its own cost structure and business model, some that are established (e.g., search), some that are emerging (e.g., chat, via search providers and others, and task automation, where companies such as UiPath, Microsoft (with Power Automate), and Automation Anywhere are taking the lead) and some that are less well defined (e.g., proactive experiences, interactive data visualization) or so-called "slow search" [10] experiences with crowdworkers employed in-house or on platforms such as Figure Eight and CrowdAI, or deep search result analysis with GenAI (e.g., the *Deep Search* feature mentioned earlier).

There are clearly many challenges that need to be considered and addressed in developing panmodal experiences (e.g., representing tasks across modalities, understanding which modalities are available and applicable for the current task, communicating relevant modality affordances to users, making transitions maximally seamless). We also must understand the balance between AI-based augmentation and automation, appreciate the learning, critical thinking, and satisfaction benefits to people from doing jobs manually, and only fully offload tasks to AI systems when we can be confident of task success and when users really do not want to do the tasks themselves (e.g., laborious chores not joyful creative acts). As with all applications of AI at scale, there are important practical considerations including safety, security, and privacy (e.g., what personal data (if any) is used for modality suggestion, how is personal data shared between modalities). Panmodal systems and experiences must be private and secure by design. As one way to do this, we could architect them with a hybrid AI approach that is similar to offerings from companies such as Google and Apple, with AI workloads split between on-device and cloud per privacy and security considerations. For example, Apple's new Private Cloud Compute, where complex and cross-modal user requests (prompts) can be encrypted and handled via stateless computation on large, cloud-based foundation models, while personal user data remains on the device and even local to the modality, would help ensure that AI inferences can safely transcend modalities. Smaller, on-device foundation models such as Phi from Microsoft and Gemini Nano from Google can also ensure that private and secure AI processing within or between modalities is feasible with less privacy and security risk. We must pay close attention to these types of issues during the design of panmodal experiences and not view the issues as an afterthought.

Only by prioritizing user trust can we build “better together” panmodal experiences that unite modalities safely and securely to help us all tackle and complete our information tasks more effectively.

REFERENCES

[1] Christopher Ahlberg, Christopher Williamson, and Ben Shneiderman. 1992. Dynamic queries for information exploration: An implementation and evaluation. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 619–626.

[2] W Bruce Croft and Roger H Thompson. 1987. I3R: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science* 38, 6 (1987), 389–404.

[3] Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining*. 1244–1247.

[4] George D Montanez, Ryen W White, and Xiao Huang. 2014. Cross-device search. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*. 1669–1678.

[5] Robert N Oddy. 1977. Information retrieval through man-machine dialogue. *Journal of Documentation* 33, 1 (1977), 1–14.

[6] Sharon Oviatt. 2007. Multimodal interfaces. *The Human-Computer Interaction Handbook* (2007), 439–458.

[7] Bo Qiao, Liqun Li, Xu Zhang, Shilin He, Yu Kang, Chaoyun Zhang, Fangkai Yang, Hang Dong, Jue Zhang, Lu Wang, et al. 2023. Taskweaver: A code-first agent framework. *arXiv preprint arXiv:2311.17541* (2023).

[8] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* 36 (2024).

[9] Ben Shneiderman. 2022. *Human-centered AI.* Oxford University Press.

[10] Jaime Teevan, Kevyn Collins-Thompson, Ryen W White, and Susan Dumais. 2014. Slow search. *Commun. ACM* 57, 8 (2014), 36–38.

[11] Ryen W White. 2024. Advancing the search frontier with AI agents. *Commun. ACM* (2024).

[12] Ryen W White, Adam Fourney, Allen Herring, Paul N Bennett, Nirupama Chandrasekaran, Robert Sim, Elnaz Nouri, and Mark J Encarnación. 2019. Multi-device digital assistance. *Commun. ACM* 62, 10 (2019), 28–31.

[13] Ryen W White, Matthew Richardson, Mikhail Bilenko, and Allison P Heath. 2008. Enhancing web search by promoting multiple search engine use. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 43–50.

[14] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155* (2023).

[15] Hamed Zamani, Johanne R Trippas, Jeff Dalton, Filip Radlinski, et al. 2023. Conversational information seeking. *Foundations and Trends® in Information Retrieval* 17, 3-4 (2023), 244–456.