User-Centric Evaluation of Query Suggestions

MSc UX Engineering
Academic Project 2023
Anyu Wang

1/Abstract

As a search assistant, query suggestions have the potential to improve search efficiency, enhance user experience, etc. Academic search is a complex and difficult task and query suggestions can address some of its challenges. The purpose of this study is to investigate the extent to which query suggestions can enhance academic searches conducted by students. Using a user-centric evaluation method, the performance of four sources of query suggestions was measured in terms of both the number of valid terms and the selection rate. The results show that query suggestions from Wordvectors outperform the others on both quantity and quality, and perform excellently under multiple topics.

2/Background

Query suggestions can improve the efficiency of academic searches, but there is a lack of user-centric evaluation of query suggestions from different sources.

Query Suggestions

Query suggestions, commonly seen in search engines, help users by offering related search terms. This makes searching faster and less stressful. However, not all suggestions improve search results. Some techniques, like adding related terms to a search, can help, but the best methods for improving searches are still debated.

Academic Search

Academic searching is a complex behaviour and varies depending on the background of the searcher. Most people will overestimate their level of academic searching. Too few keywords in a query can lead to poor search results. The high level query suggestions provided by query expansion techniques can eliminate the above issues.

System-centric VS User-centric

The common evaluation methods and corresponding metrics are accuracy, precision, recall, MRR, nDCG, etc. These are all system-centric evaluation methods. The sample sizes are very large, with the largest having data collection spanning several years. In terms of evaluating applications of query suggestions, many user-centric approaches aim to evaluate interaction and interface design. I have learned from and refined some of the user-centric terms evaluation methodology.

5/Conclusion

After comprehensive analyses, Wordvetors was found to be outstanding in terms of the quantity and quality of query suggestions, as well as applicability, and performed well under keywords across a wide range of topics.There were significant differences between Wordvectors' performance and the other three sources. But the remaining three, MeSH, PubMed, and WebIsA, did not show significant differences between each other. MeSH performs well for terms in specific domains, but generally does not work well for academic searches across disciplines. In the future, it is possible to deeply explore the reasons for the variability of the different sources and how they can be combined with actual searches to achieve the best results.

3/Methodology

Participants were asked to complete 20 academic search tasks on different topics, and they were free to search with the help of query suggestions until they were satisfied with the results. The performance of the sources was evaluated by the number of query suggestions generated and the selection rate of the participants.

Sources

Source1
MeSH

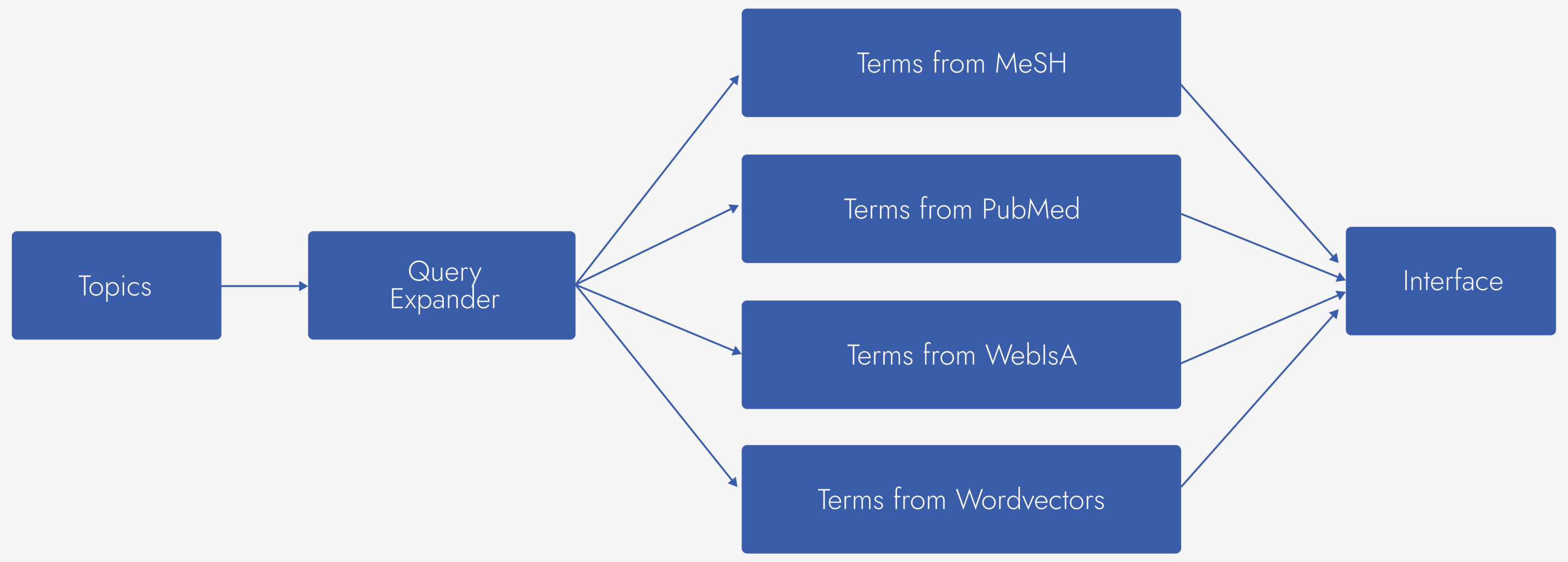
Source2
PubMed

Source3
WebIsA

Source4
Wordvectors

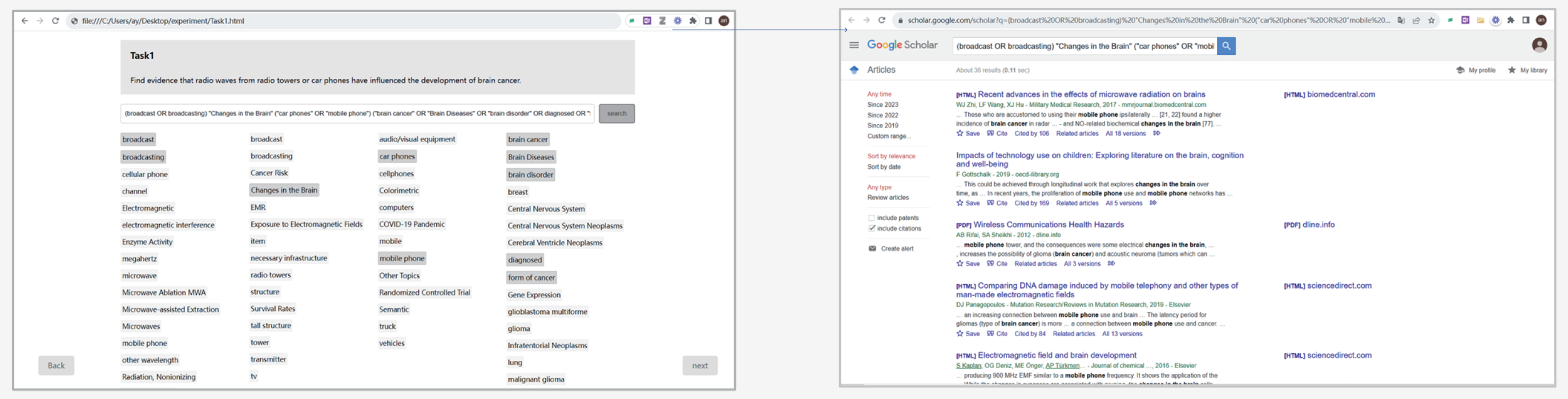
Terms Generation

I generated query suggestions from different sources based on keywords on the same topic and put them on the interface where the sources were not visible to the participants. Participants conducted academic searches with the help of query suggestions provided by the study. The performance of the sources was evaluated by the number of query suggestions generated and the selection rate of the participants.



Interface

The experiment has 22 pages. Participants performed the 20 main tasks after passing the example task on the first page. Each task was on a single page. Participants clicked to select terms that were combined by Boolean operators and displayed in the search bar. Through the Google Scholar API, the content of the search bar is searched with the Google Scholar search engine.


file:///C:/Users/ay/Desktop/experiment/Task1.html

Task1

Find evidence that radio waves from radio towers or car phones have influenced the development of brain cancer.

(broadcast OR broadcasting) "Changes in the Brain" ("car phones" OR "mobile phone") ("brain cancer" OR "Brain Diseases" OR "brain disorder" OR diagnosed OR "

scholar.google.com/scholar?q=(broadcast%20OR%20broadcasting)%20"Changes%20in%20the%20Brain"%20("car%20phones"%20OR%20"mobile%20...

Google Scholar (broadcast OR broadcasting) "Changes in the Brain" ("car phones" OR "mobi

Articles — About 36 results (0.11 sec)

[HTML] Recent advances in the effects of microwave radiation on brains
WJ Zhi, LF Wang, XJ Hu - Military Medical Research, 2017 - mmrjournal.biomedcentral.com
… Those who are accustomed to using their **mobile phone** ipsilaterally … [21, 22] found a higher incidence of **brain cancer** in radar … - and NO-related biochemical **changes in the brain** [77]. …
Save Cite Cited by 106 Related articles All 18 versions
[HTML] biomedcentral.com

Impacts of technology use on children: Exploring literature on the brain, cognition and well-being
F Gottschalk - 2019 - oecd-ilibrary.org
… This could be achieved through longitudinal work that explores **changes in the brain** over time, as … In recent years, the proliferation of **mobile phone** use and **mobile phone** networks has …
Save Cite Cited by 169 Related articles All 5 versions

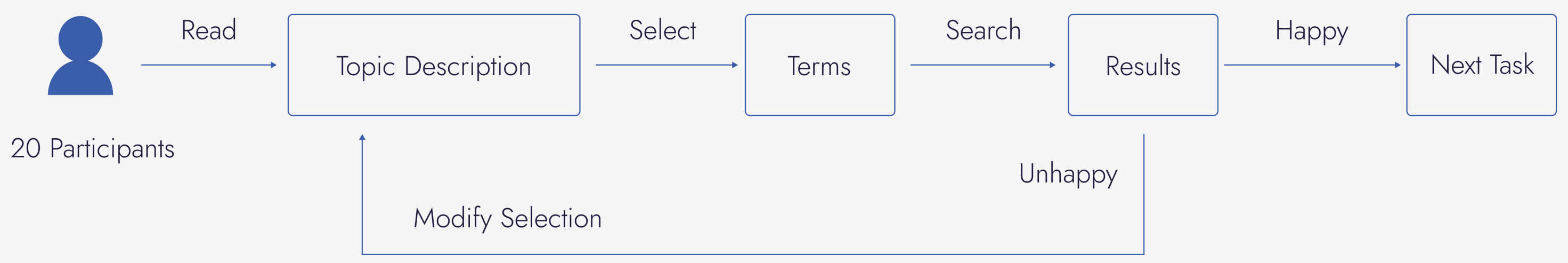
[PDF] Wireless Communications Health Hazards
AB Rifai, SA Sheikhi - 2012 - dline.info
… **mobile phone** tower, and the consequences were some electrical **changes in the brain**, … , increases the possibility of glioma (**brain cancer**) and acoustic neuroma (tumors which can …
Save Cite Related articles All 3 versions
[PDF] dline.info

[HTML] Comparing DNA damage induced by mobile telephony and other types of man-made electromagnetic fields
DJ Panagopoulos - Mutation Research/Reviews in Mutation Research, 2019 - Elsevier
… an increasing connection between **mobile phone** use and brain … The latency period for gliomas (type of **brain cancer**) is more … a connection between **mobile phone** use and cancer. …
Save Cite Cited by 84 Related articles All 13 versions
[HTML] sciencedirect.com

[HTML] Electromagnetic field and brain development
S Kaplan, OG Deniz, ME Önger, AP Türkmen… - Journal of chemical …, 2016 - Elsevier
… producing 900 MHz EMF similar to a **mobile phone** frequency. It shows the application of the
[HTML] sciencedirect.com

Operations

Academic searches are usually multiple queries to obtain a single piece of expected information. Therefore this experiment allows for multiple searches to obtain more accurate results.

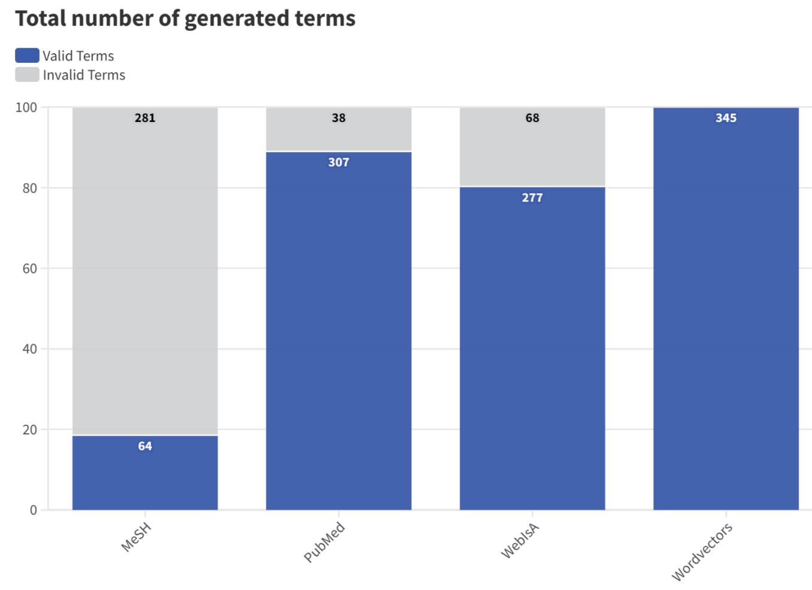


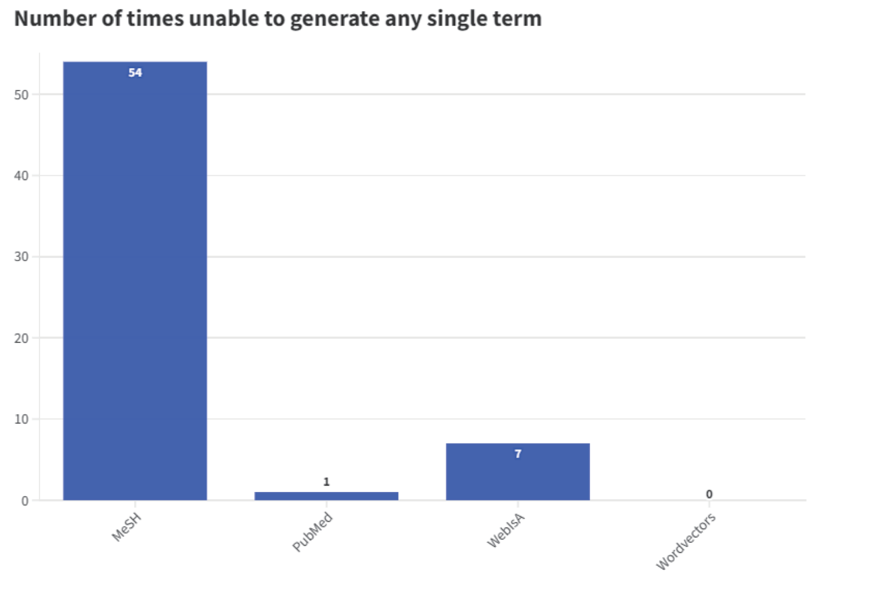
4/Results

The number of valid terms and the selection rate show the performance of the different sources from the aspect of quantity and quality of query suggestions. The more valid terms and the higher the selection rate, the better the performance.

Valid Terms

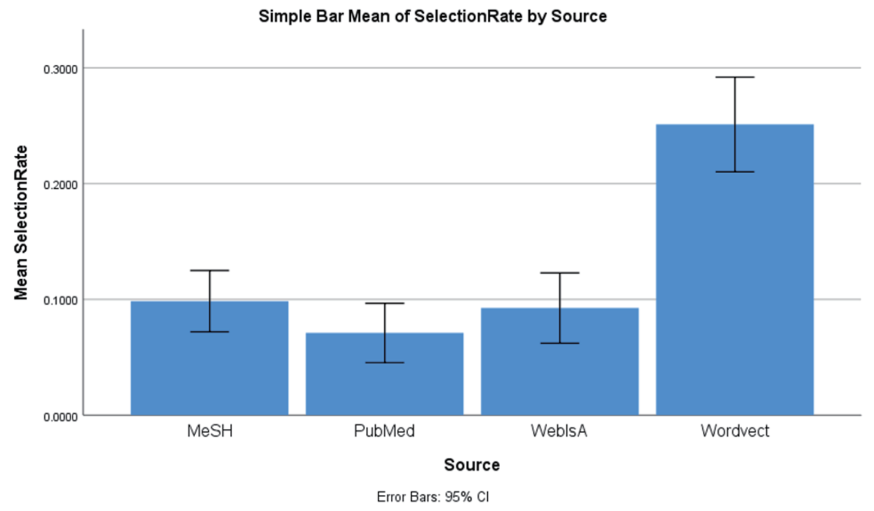
The number of valid terms objectively reflects the performance of the source.Wordvectors performed best, generating a sufficiently high number of query suggestions for each keyword. Out of a total of 69 keywords, 54 were not able to get query suggestions through MeSH.The number of valid terms for PubMed and WebIsA was close and in an acceptable range.


Total number of generated terms

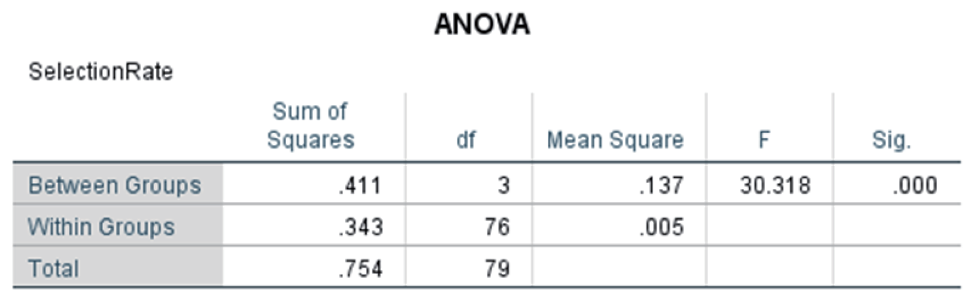

Number of times unable to generate any single term

Selection Rate

The selection rate reflects the quality of valid terms. A higher selection rate indicates a higher quality of terms from the source. As can be seen from the figure, the selection rate of Wordvectors is significantly different from that of all three other sources, and the average selection rate is much higher. Although MeSH generates fewer terms, it has the second highest selection rate following Wordvectors. But there is no significant difference between PubMed, MeSH and WebIsA.

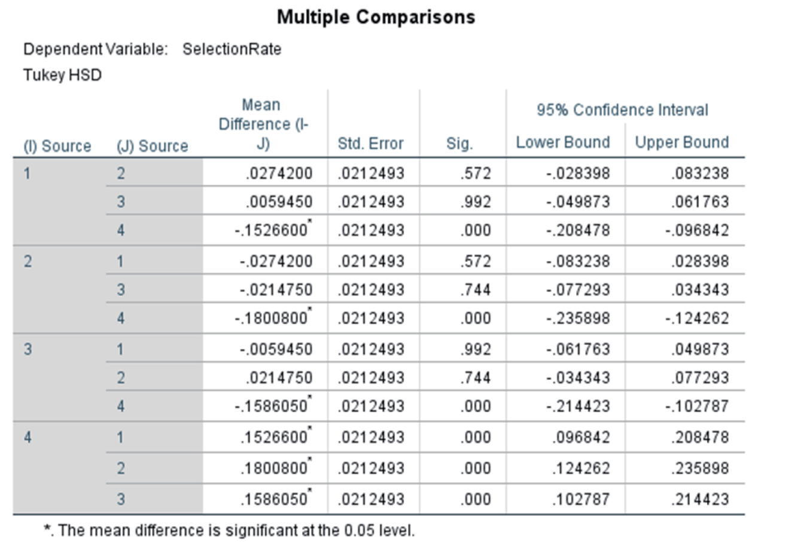

Simple Bar Mean of SelectionRate by Source

Error Bars: 95% CI


ANOVA

SelectionRate

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.411	3	.137	30.318	.000
Within Groups	.343	76	.005		
Total	.754	79			

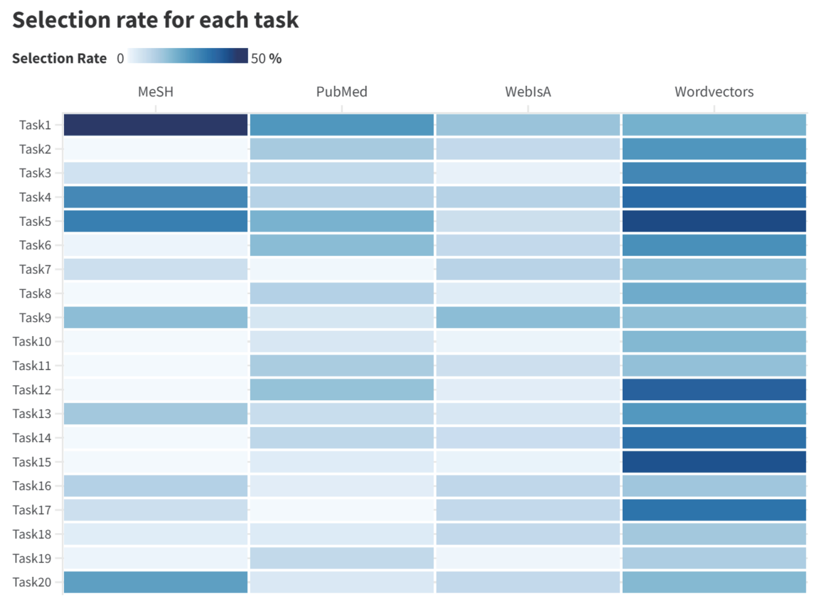

Multiple Comparisons

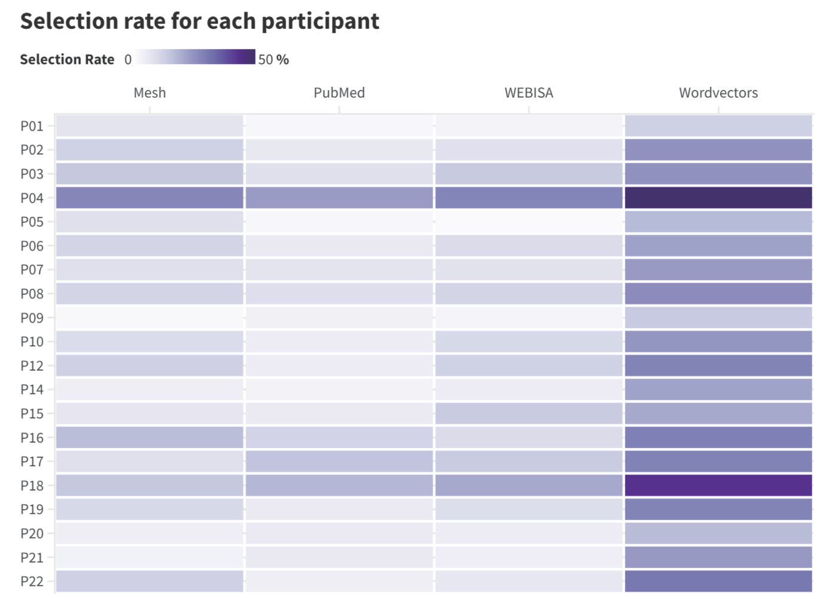
Dependent Variable: SelectionRate
Tukey HSD

(I) Source	(J) Source	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	.0274200	.0212493	.572	-.028398	.083238
	3	.0059450	.0212493	.992	-.049873	.061763
	4	-.1526600*	.0212493	.000	-.208478	-.096842
2	1	-.0274200	.0212493	.572	-.083238	.028398
	3	-.0214750	.0212493	.744	-.077293	.034343
	4	-.1800800*	.0212493	.000	-.235898	-.124262
3	1	-.0059450	.0212493	.992	-.061763	.049873
	2	.0214750	.0212493	.744	-.034343	.077293
	4	-.1586050*	.0212493	.000	-.214423	-.102787
4	1	.1526600*	.0212493	.000	.096842	.208478
	2	.1800800*	.0212493	.000	.124262	.235898
	3	.1586050*	.0212493	.000	.102787	.214423

*. The mean difference is significant at the 0.05 level.

Wordvectors had high selection rates in general and could be used in a wide range of topics. PubMed and WebIsA also had similar and average overall performance. MeSH had very high selection rates in a few tasks and zero selection rates in nearly half of the tasks. Furthermore, MeSH performed well in specific specialised terms, but poorly in generic ones. A few participants chose more terms and spent more time on them, but overall almost all participants tended to select terms from Wordvectors.


Selection rate for each task


Selection rate for each participant