

## 论文摘要

随着互联网上虚假新闻的兴起, 对其进行有效甄别成为社会的迫切需要。在判断新闻真伪时, 用户会使用搜索引擎获取相关信息。分析和理解用户如何利用互联网信息判断新闻真伪, 有助于设计更好的具有可解释性的虚假新闻检测算法。该文通过用户实验, 收集了用户在使用搜索引擎判断新闻真假时的行为日志。借助日志, 该文从真假判断是否变化、真假判断是否正确、真假判断是否正确三个角度对用户的相关行为指标进行了分析。结果显示, 用户在不同场景下的搜索行为存在较大差异, 表明用户对新闻真假的认知影响了其在使用搜索引擎时的行为。

## 研究问题

1. 当用户从搜索引擎上获取的信息影响了其判断时, 搜索行为会如何变化?
2. 当用户在使用搜索引擎对缺乏先验知识的难以直接判断真假的新闻进行真假判断时, 搜索行为会如何变化?
3. 当用户从搜索引擎上没有获取到有用信息时, 其搜索行为具有哪些特征?
4. 当用户从搜索引擎上获取到了错误的信息时, 其搜索行为具有哪些特征?

## 实验流程与结果

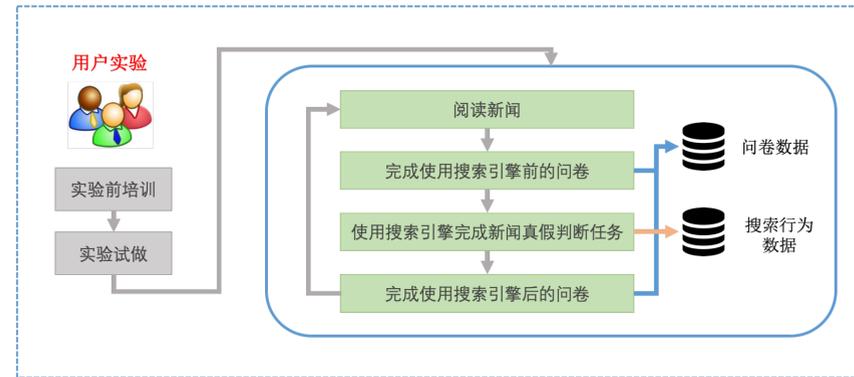


表 1 问卷收集的信息

问卷	收集的信息	可回答的选项
使用搜索引擎前的问卷	新闻的真假标注 (使用搜索引擎前)	0=假, 1=不确定, 2=真
使用搜索引擎后的问卷	新闻的真假标注 (使用搜索引擎后)	0=假, 1=不确定, 2=真
	改正结果	如果在使用搜索引擎后判断新闻是假, 需要用自然语言描述新闻的改正结果
	点击的文档是否有用	0=无用, 1=有用

表 2 浏览器扩展程序记录和计算出的搜索行为指标

搜索行为指标	英文记号	含义
查询词数量	query_num	判断某条新闻真假总共输入的查询词数量
总搜索页面停留时间	total_SERP_time	判断某条新闻真假在搜索页面上停留的总时间
每个查询词下平均搜索页面停留时间	avg_SERP_time	判断某条新闻真假平均在每个查询词下的搜索页面停留时间
总文档点击数	total_click_count	判断某条新闻真假总文档点击数
每个查询词下平均文档点击数	avg_click_count	判断某条新闻真假平均在每个查询词下的文档点击数
点击文档中有用文档比例	useful_click_ratio	判断某条新闻真假点击文档中有用文档的比例

表 3 用户各搜索行为指标平均值 (真假判断变化 vs 不变)

	整体	真假判断		p value
		变化	不变	
新闻任务数量	207	93	114	--
query_num	1.99	1.95	2.03	0.6773
total_SERP_time(s)	44.2	39.7	48.0	0.0835*
avg_SERP_time(s)	22.7	20.5	24.5	0.0456**
total_click_count	2.98	3.38	2.65	0.0046***
avg_click_count	1.74	2.01	1.53	0.0015***
useful_click_ratio	0.724	0.735	0.716	0.6934

表 5 用户各搜索行为指标平均值 (使用搜索引擎后真假判断确定 vs 不确定)

	整体	使用搜索引擎后真假判断		p value
		确定	不确定	
新闻任务数量	207	186	21	--
query_num	1.99	1.92	2.62	0.0264**
total_SERP_time(s)	44.2	41.4	69.4	0.0003***
avg_SERP_time(s)	22.7	22.2	26.9	0.1621
total_click_count	2.98	2.96	3.10	0.7555
avg_click_count	1.74	1.79	1.35	0.0821*
useful_click_ratio	0.724	0.739	0.595	0.0713*

表 4 用户各搜索行为指标平均值 (使用搜索引擎前真假判断确定 vs 不确定)

	整体	使用搜索引擎前真假判断		p value
		确定	不确定	
新闻任务数量	207	127	80	--
query_num	1.99	1.95	2.05	0.6208
total_SERP_time(s)	44.2	45.7	41.9	0.4361
avg_SERP_time(s)	22.7	24.2	20.3	0.0630*
total_click_count	2.98	2.77	3.30	0.0448**
avg_click_count	1.74	1.65	1.89	0.1309
useful_click_ratio	0.724	0.725	0.723	0.9646

表 6 用户各搜索行为指标平均值 (前正确后错误 vs 前错误后正确)

	整体	真假判断		p value
		前正确后错误	前错误后正确	
新闻任务数量	207	12	67	--
query_num	1.99	2.08	1.99	0.8228
total_SERP_time(s)	44.2	52.0	38.4	0.1820
avg_SERP_time(s)	22.7	24.9	19.4	0.0970*
total_click_count	2.98	3.50	3.43	0.9152
avg_click_count	1.74	2.08	1.99	0.7763
useful_click_ratio	0.724	0.708	0.740	0.7599

## 论文结论

1. 当从搜索引擎上获取的信息影响了用户的真假判断时, 用户会在搜索页面上停留更短的时间, 并点击更多的文档。这可能是因为, 用户对新闻的真假判断发生变化时需要较多的信息支持, 用户需要进行更多点击来获取信息, 在搜索页面上停留的时间就会变短。
2. 当在使用搜索引擎对缺乏先验知识的难以直接判断真假的新闻进行真假判断时, 用户在每个查询下的搜索页面停留时间会更短, 总的文档点击数会更多。这可能是因为, 对不熟悉的内容, 用户往往会倾向于点击更多的文档, 以获取对该内容的基本认识。自然地, 用户在搜索页面上停留的时间就会比较短。
3. 当从搜索引擎上没有获取到有用信息时, 用户在搜索页面上停留的时间会更长, 在每个查询下的点击会更少, 并且会更多地输入新的查询, 其点击文档中有用文档的比例也会更少。这可能因为, 当用户输入查询词后, 会发现搜索页面上展示的大多是无关的信息, 因此用户会在搜索页面上停留较长的时间去决定点击哪个文档, 自然在每个查询词下的平均点击数就会比较少。当用户发现这个查询下搜索不到有用的信息时, 就会输入一个新的查询, 因此查询数量会比较多。另外, 由于用户没有获取到有用的信息, 其点击文档中有用文档的比例自然也会比较低。
4. 当从搜索引擎上获取到错误的信息时, 用户在每个查询下的平均搜索页面停留时间会更长。这可能是因为用户无法有效分辨获取信息的真伪, 导致其做出错误的判断。由于信息纷杂, 用户需要在搜索页面上停留更多时间, 来分辨哪些文档是值得点击的。