TREO
Technology, Research, Education, Opinion

Understanding Long tail effect in Search Engines : A Study of User Queries by evaluating Content and Sequence measures

Aarushi Jain (f17aarushij@iimidr.ac.in); Rajhans Mishra (rajhansm@iimidr.ac.in)
Indian Institute of Management Indore; Indian Institute of Management Indore

The transformation of mass market into millions of small niche market is known as the long tail concept. Each of the niche markets maybe small but the combination of various niches creates greater volume of business in the traditional mass market success (Chris Anderson,2004). The evolution of the long tail came from the e-commerce industry where the 'one-shop-stop' solution such as Amazon.com, Netflix and iTunes have created the magnitude of plethora of items than the brick and mortar services resulting in the long tail market. Similarly, the presence of search engine complexity and the increase in the computing power of search engines the users have started practicing the long tail by writing longer queries in the web. This helps the users to reach to a specific and niche content in information retrieval process. The availability of homogeneous information which is required by the users is not easy as the size of the web has increased and mammoth information is present on the web. Therefore, the users are writing longer queries to reach out to the niche and specific information.

The number of search engines are increasing due to the presence of thousands of web services available for the users. Due to this, the nature of the search engines is changing due to the increase dynamism in the web (Vaughan, 2004). The reason for the changing nature of the web is due to the increase in dynamism in the web (Tourani & Danesh, 2013) and the users use long phrases for searching the content on the web (Lau et al, 1999 & Phan et al, 2007). Earlier, the search queries for Information Retrieval on the web were short. There have been several studies done to study query logs from major search engines (Spink & Jansen, 2004). The study concludes that the average query length on the web done are only two to three terms long and the queries used for retrieval of information on the web covers various topics. However, the longer queries are discouraged in the web search engines. There is a filtering process for long keywords to be included in the search results. To match the search results of the user's queries, the query terms must be included in either the pages of the results or must be in the anchor texts of the links referencing it. For instance, "What is", "How to" were not earlier included in the search results but now these words are replaced by the synonyms in the search results. Till date, for the longer queries, the web search engines perform poorly and do not provide the match to the user's queries as while doing the web search, the different search engine gives different results which implies that there are variations in the outputs of the results provided by different search engines. From the user perspective, they are not only influenced by the content of the search results but also influenced by the sequence of the search results (Joachims et al, 2007; Bar-Ilan et al, 2009). Therefore, in this paper, we are exploring the variations in the search engine outputs using the content of search results retrieved and also the display of the order of the search results i.e., sequence is retrieved. The comparison of three different search engines: Google, Yahoo and Bing are done in terms of the content & sequence similarity measures for longer queries. We are comparing the URLs for the identical queries retrieved from the three search engines. For comparing the content similarity of the URL's, Jaccard's similarity is computed and for comparing the sequence similarity of the URL's, the sequence mismatch of the search results in the three search engines is evaluated. This paper proposes a methodology for evaluating the performance of the search engines on the aforementioned measures by conducting experiments. This approach can be useful for any online services such as e-commerce companies, music companies, tourism industry for targeting the customers who use these search engines for online shopping and digital marketing.

Keywords- Long tail effect, user queries, search engines performance, Jaccard's measure, sequence similarity.

References

Anderson, C., & Andersson, M. P. (2004). Long tail.

Bar-Ilan, J., Keenoy, K., Levene, M., & Yaari, E. (2009). Presentation bias is significant in determining user preference for search results—A user study. Journal of the American Society for Information Science and Technology, 60(1), 135-149

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. ACM Transactions on Information Systems (TOIS), 25(2), 7.

Lau, T., & Horvitz, E. (1999). Patterns of search: analyzing and modeling web query refinement. In UM99 user modeling (pp. 119-128). Springer, Vienna.

Phan, N., Bailey, P., & Wilkinson, R. (2007, July). Understanding the relationship of information need specificity to search query length. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 709-710). ACM

Spink, A., & Jansen, B. J. (2004). A study of web search trends. Webology, 1(2), 4.

Tourani, A., & Danesh, A. S. (2013). Using Exclusive Web Crawlers to Store Better Results in Search Engines' Database. *arXiv preprint arXiv:1305.2686*.

Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. *Information Processing & Management*, *40*(4), 677-691.