queries on a topic not covered by the page). Over the past few years, many heuristics have been proposed to identify spam web pages and sites, see for example the series of AIRweb workshops [7]. The problem of identifying web spam can be framed as a classification problem, and there are many well-known classification approaches (e.g., decision trees, Bayesian classifiers, support vector machines).

5. Deep Web Crawling

The deep web crawling problem is closely related to the problem known as federated search or distributed information retrieval, in which a mediator forwards user queries to multiple searchable collections, and combines the results before presenting them to the user.

5.1 Problem Overview

Deep web crawling has three steps:
(1) **Locate deep web content sources.** A human or crawler must identify web sites containing form interfaces that lead to deep web content.

(2) **Select relevant sources.** For a scoped deep web crawling task (e.g., crawling medical articles), one must select a relevant subset of the available content sources. In the unstructured case this problem is known as database or resource selection. The first step in resource selection is to model the content available at a particular deep web site,e.g., using query-based sampling.

(3) **Extract underlying content.** Finally, a crawler must extract the content lying behind the form interfaces of the selected content sources

6. Future Work- Conclusion

As this study indicates, crawling is a well-studied problem. However, there are at least as many open questions as there are resolved ones. Even in the material we have covered, the reader has likely noticed many open issues, including:

1) **Parameter tuning.** Many of the crawl ordering policies rely on carefully tuned parameters, with little insight or science into how best to choose the parameters. For example, what is the optimal level of greediness for a scoped crawler
2) **Retiring unwanted pages.** Given finite storage capacity, in practice crawlers discard or retire low-quality and spam pages from their collections, to make room for superior pages.
3) However, we are not aware of any literature that explicitly studies retirement policies. There is also the issue of how much metadata to retain about retired pages, to avoid accidentally
4) rediscovering them, and to help assess the quality of related pages (e.g., pages on the same site, or pages linking to or linked from the retired page).
5) **Holistic crawl ordering.** Whereas much attention has been paid to various crawl ordering sub-problems (e.g., prioritizing the crawl order of new pages, refreshing content from old pages, revisiting pages to discover new links), there is little work on how to integrate the disparate approaches into a unified strategy.
6) **Deep web.** Clearly, the science and practice of deep web crawling is in its infancy. There are also several nearly untouched directions.
7) **Crawling scripts.** Increasingly, web sites employ scripts (e.g., JavaScript, AJAX) to generate content and links on the fly. Almost no attention has been paid to whether or how to crawl these sites.
8) **Personalized content.**Web sites often customize their content to individual users, e.g., Amazon gives personalized recommendations based on a user's browsing and purchasing
9) patterns. It is not clear how crawlers should treat such sites, e.g., emulating a generic user versus attempting to specialize the crawl based on different user profiles. A search engine that aims to personalize search results may wish to push some degree of personalization into the crawler.
10) **Collaboration between content providers and crawlers.** Crawling is a pull mechanism for discovering and acquiring content. Modern commercial crawlers employ a hybrid of push and pull, but there is little academic study of this practice and the issues involved.

References

[1] S. Abiteboul, M. Preda, and G. Cobena, "Adaptive on-line page importance computation," in Proceedings of the 12th International World Wide Web Conference, 2003.
[2] E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas, "The web changes everything: Understanding the dynamics of web content," in Proceedings of the 2nd International Conference on Web Search and Data Mining, 2009.
[3] Advanced Triage (medical term), http://en.wikipedia.org/wiki/Triage# Advanced triage.
[4] A. Agarwal, H. S. Koppula, K. P. Leela, K. P. Chitrapura, S. Garg, P. K. GM, C. Haty, A. Roy, and A. Sasturkar, "URL normalization for de-duplication of web pages," in Proceedings of the 18th Conference on Information and Knowledge Management, 2009.
[5] C. C. Aggarwal, F. Al-Garawi, and P. S. Yu, "Intelligent crawling on the world wide web with arbitrary predicates," in Proceedings of the 10th International World Wide Web Conference, 2001.
[6] D. Ahlers and S. Boll, "Adaptive geospatially focused crawling," in Proceedings of the 18th Conference on Information and Knowledge Management, 2009.
[7] International Workshop Series on Adversarial Information Retrieval on the Web, 2005–

Author Profile

Pallavi received the B.TECH degree in information technology and M.TECH. degree in Computer Science and Engineering from Maharshi Dayanand University in 2013 and 2015,respectively.