Retrieval-Augmented Generation (RAG) has been shown to improve knowledge capabilities and alleviate the hallucination problem of LLMs. The Web is a major source of external knowledge used in RAG systems, and many commercial systems such as ChatGPT and Perplexity have used Web search engines as their major retrieval systems. Typically, such RAG systems retrieve search results, download HTML sources of the results, and then extract plain texts from the HTML sources. Plain text documents or chunks are fed into the LLMs to augment the generation. However, much of the structural and semantic information inherent in HTML, such as headings and table structures, is lost during this plain-text-based RAG process. To alleviate this problem, we propose HtmlRAG, which uses HTML instead of plain text as the format of retrieved knowledge in RAG. We believe HTML is better than plain text in modeling knowledge in external documents, and most LLMs possess robust capacities to understand HTML. However, utilizing HTML presents new challenges. HTML contains additional content such as tags, JavaScript, and CSS specifications, which bring extra input tokens and noise to the RAG system. To address this issue, we propose HTML cleaning, compression, and pruning strategies, to shorten the HTML while minimizing the loss of information. Specifically, we design a two-step block-tree-based pruning method that prunes useless HTML blocks and keeps only the relevant part of the HTML. Experiments on six QA datasets confirm the superiority of using HTML in RAG systems.