

All You Wanted To Know About Web Information Extraction

Ever wondered how the directories especially the search engine-type listings get the exact details about the webpages with their aggregated sites database? How do they gather information from the webpages so easily? Well, if you think they deploy some web information extraction strategies to derive the right information, you have nailed it!!! Web Information Extraction

Web information extraction refers to the manual activity of gathering information on content or Meta elements to compile a list for public viewing. Most directories exist not for earning money but for the purpose of offering convenient source of web information to the internet users. It is one of the most important and fundamental activities that search engines usually take the help of in order to offer information on the existing websites on the World Wide Web. Hence, so long as new websites keep popping up, search engines would always expand the web information extraction on them. Usually, search engine directories utilize common crawling or listing strategies to obtain the exact information from the source of the webpages themselves. There are two very popular procedure of obtaining information on webpage details such as manual data extraction and entry, and robot crawling. These techniques target the Meta tags that contain the title, description and link information. This is how top search engines and directories such as Google, Yahoo and GoGuides.org obtain information about all the websites that are registered within their categories.

Manual Web Information Extraction

The manual process consists of scanning of websites for title and description which are basically carried out by data entry personnel. After this process, they link the sites through different categories depending on the use and relevance factors of the websites. Sometime, the sites are checked for quality of the content and visual items. When these procedures are over, the websites are left for the public viewing. Understandably, manual directory information extraction is a dreary job; however, it is very effective as they offer highly unique data listings. Some human edited directories such as DMOZ look for volunteer web editors to assist in developing their directory. More often than not, these editors often create very original site descriptions and this is precisely why most search engines crawl the information from DMOZ itself because of the authenticity of the information available on it. Automated Web Information Extraction

When it comes to the automated procedure, the techniques are almost similar with those used in the manual process. Search engine crawlers look for given fields such as Meta titles and description from the page source and then compile them in an order depending on their relevance and purpose with respect to the sites. The crawlers then collate all the information obtained and display them on the site directory. The entire process is carried out automatically by coding applications that go from one site to another to list the necessary details. However, this automatic automotive procedure has its own limitations. For a starter, this type of strategy produces repetitive site descriptions and there is also a chance that it could lead to fallacious descriptions. As a matter of fact, robot crawlers can only extract web information and, can't edit it to rectify the data form. They are incapable of deciding if the information entered in those sites is relevant or just supplementary to the category.

About the Author

Maneet Puri heads LeXolution IT Services (LIT), a renowned web and KPO solution provider in India. He is an expert of web based applications which include web design, web development, and [website maintenance services](#). Besides, he has also successfully added KPO to his company. He has efficiently handled many off-shore projects related to [web design & web development](#) for his overseas clients. If you want more information on web application services offered by LIT, visit www.lexolutionit.com or visit the blog www.all-that-web-demans.blogspot.com.

Source: <http://www.serverforever.com>