Abstract

Search engine is an application which capable to index URLs from internet so it have large web full description *database*. This "text on images search engine" perform more specific searching, it is text appeared on image. This machine contain 3 subsystems, they are crawling, text recognition, and searching it self.

Crawling process has job to crawl URLs, find, and download image on every page. This job start with a feed URL given by admin. In other side, *text recognition* process work parallel with this crawling process. The images mined, will be processed sequencially. The main duty of recognition process is to ectract text from image. And then, these text divide into *term*s, saved, and indexed on *database*. Part of system that have direct interaction with *user* is search subsystem. In this application, *user* has to enter keywords of searching on the text input form, and system will find *term*s in *database* that considere to the keywords. Finally, the result images will be showed to *user*.

Overall, this system provides search engine with specific features, *similar match* that search similar term with user query input, and *exact match* that search same term with user query input. These options available because of the text extraction results weren't always exactly same with the real text on image.

Keywords: *crawling*, *text recognition*, *searching, term*