

Technical Disclosure Commons

Defensive Publications Series

February 05, 2019

In-page search over text and images in a web page

Matthew Sharifi

Follow this and additional works at: https://www.tdcommons.org/dpubs_series

Recommended Citation

Sharifi, Matthew, "In-page search over text and images in a web page", Technical Disclosure Commons, (February 05, 2019)
https://www.tdcommons.org/dpubs_series/1931



This work is licensed under a [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/).

This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.

In-page search over text and images in a web page

ABSTRACT

Browser applications provide users with the ability to search within a webpage. Such searches are limited to textual content of the page and do not take into account content depicted as images within the webpage. This disclosure describes techniques to enable search to be performed over the entire webpage, including image content. With user permission, images on a webpage are analyzed to determine presence of text within the image. If text is detected, OCR techniques are applied to recognize characters in the text. When the user enters a search query for searching within the page, the text content of the webpage and text obtained via OCR from images within the webpage are compared with the search query. Matching terms are highlighted to the user, with the image content highlighting based on a bounding box for the matching term.

KEYWORDS

- webpage
- optical character recognition (OCR)
- alt text
- alt tag
- in-page search
- browser

BACKGROUND

Browser applications provide users with the ability to search within a webpage. Such searches are limited to textual content of the page and do not take into account content depicted as images within the webpage. For example, when a user performs a search in the browser for a

symbol within a formula that is rendered as an image, the browser does not return the formula as a matching result.

DESCRIPTION

User permission is obtained to perform automatic analysis of a webpage that the user is viewing. When the user provides permission, optical character recognition (OCR) techniques are applied to extract text from each of the images on the webpage. This recognition can either be done offline on a server (e.g., at a time the webpage is crawled or loaded) and dynamically fetched by the client, or can be performed locally on the device.

When the recognition is performed on-device, alt tag hints (e.g., that include alternate text for the image) and a lightweight text detector can be used to guide the search towards images that likely need be subjected to OCR. This is particular useful in certain devices, e.g., mobile devices, or other devices with low computing capability, since OCR over all images might not be feasible on such devices

Once text has been derived from images on the page, the standard search mechanism can be used to find both text and image matches in the order which they appear on the page. A consistent UI is used to highlight matches across text and images. In the case of image matches, the underlying bounding box obtained from the OCR engine is highlighted. The OCR process utilizes a lightweight text detector that detects the presence of text in images), followed by the OCR engine that accepts pixels as input and detects characters. For example, the OCR engine can be a deep neural network, e.g. a convolutional network.

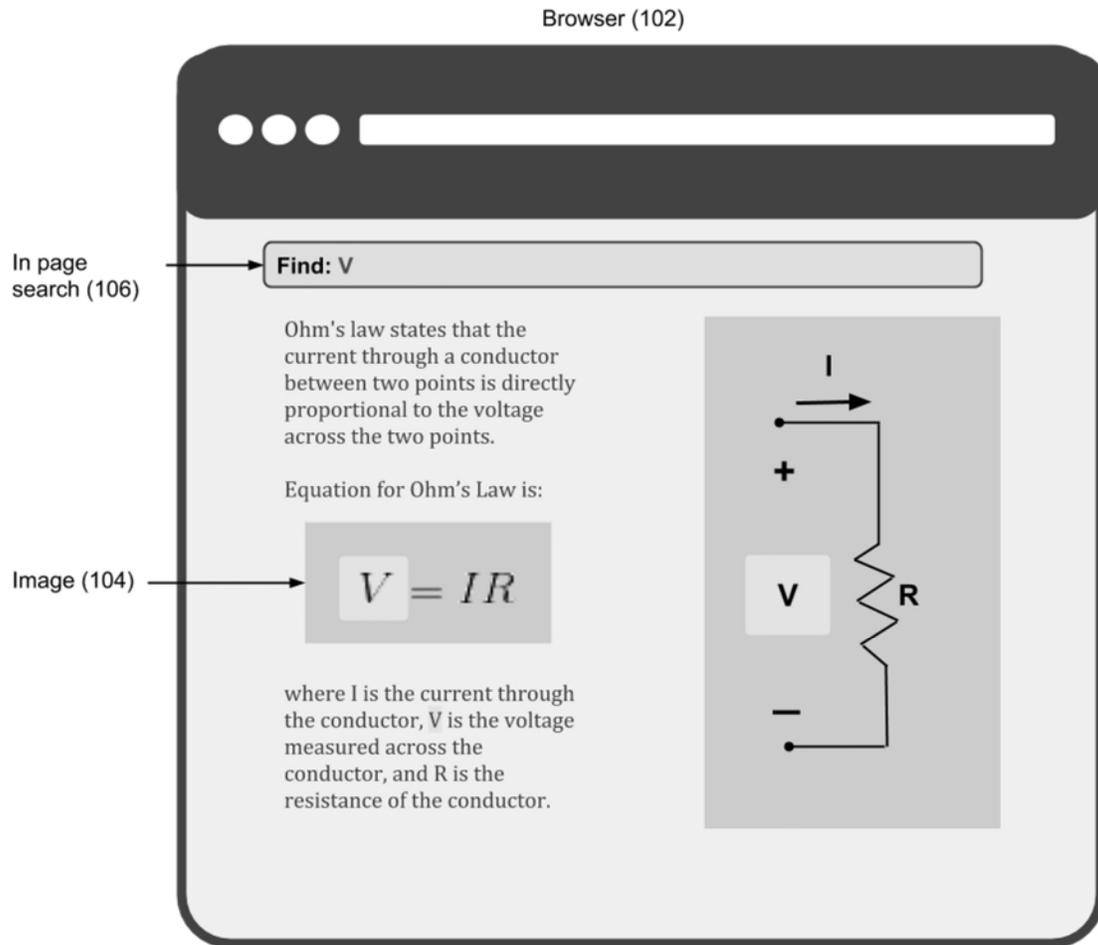


Fig. 1: In page search over text and images in a browser

Fig. 1 illustrates an example in which search in a browser application (102) is performed over the entire web page content, including text and images (e.g., image 104). In this example, a webpage related to Ohm's law is being viewed in the browser. The equation for Ohm's law (" $V=IR$ ") and an example circuit illustrating the flow of current are included in the webpage as images.

In the illustrated example, the user performs an in-page search (106), e.g., for the term "V." In response, the browser detects that "V" is present in the equation, the circuit diagram, and

in the text below the equation and highlights the text for the user. The highlighting within the image is bounded by a box, e.g., such that it includes only the search term - the letter “V.”

The described techniques are implemented with specific user permission. If the user denies permission to analyze webpage content for search over images, image analysis is not performed. Users are provided with options to turn off image-to-OCR.

Further to the descriptions above, a user is provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein enable collection of user information (e.g., information about a user’s social network, social actions or activities, profession, a user’s preferences, or a user’s current location), and if the user is sent content or communications from a server. In addition, certain data is treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user’s identity is treated so that no personally identifiable information can be determined for the user, or a user’s geographic location is generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user has control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes techniques to enable search to be performed over the entire webpage, including image content. With user permission, images on a webpage are analyzed to determine presence of text within the image. If text is detected, OCR techniques are applied to recognize characters in the text. When the user enters a search query for searching within the page, the text content of the webpage and text obtained via OCR from images within the

webpage are compared with the search query. Matching terms are highlighted to the user, with the image content highlighting based on a bounding box for the matching term.

REFERENCES

1. Jaderberg, Max, Andrea Vedaldi, and Andrew Zisserman. "Deep features for text spotting." In European conference on computer vision, pp. 512-528. Springer, Cham, 2014.