
http://researchrepository.murdoch.edu.au/13170/

This paper is posted here for your personal use. No further distribution is permitted.
A Web Assessment Approach based on Summarisation and Visualisation

Wigrai Thanadechteemapat 1, Chun Che Fung 2

School of Information Technology, Murdoch University
South Street, Murdoch, Western Australia
1 w.thanadechteemapat@murdoch.edu.au
2 l.fung@murdoch.edu.au

Abstract—The number of Web sites has noticeably increased to roughly 224 million in last ten years. This means there is a rapid growth of information on the Internet. Although search engines can help users to filter their desired information, the searched result is normally presented in the form of a very long list, and users have to visit each Web page in order to determine the appropriateness of the result. This leads to a considerable amount of time has to be spent on finding the required information. To address this issue, this paper proposes a Web assessment approach in order to provide an overview of the information on a Website using an integration of existing summarisation and visualisation techniques, which are text summarisation, tag cloud, Document Type View, and interactive features. This approach is capable to reduce the time required to identify and search for information from the Web.

I. INTRODUCTION

Information on the Internet is expanding at a rapid growth rate and no central entity is responsible for organising the information [1]. The number of Web sites has dramatically increased to approximately ten times at 224 million in between 1996 to 2009 [2]. This had led to the phenomenon of Information overloading [3]. While search engines have played an important role in assisting users to look for information on the Internet [1], such engines use their own algorithms or mechanisms based on different techniques to rank the Websites. In particular, the results are not based on the content or context from the websites. Hence, useful information or appropriate sites may be overlooked. In addition, the searched results provided by the search engines are normally an extensive long list, and users have to visit each individual link in order to determine whether the link contains the required information. These activities will require a lot of time to traverse the websites page by page.

One solution to address this problem is to provide an overview of each Web page to the users so they could assess whether the information meets their needs. Therefore, this paper proposes a web assessment approach based on summarisation and visualisation techniques. This approach is believed to be able to reduce the time required by the users to identify their required information on the Web.

This paper starts with an introduction and describes the aims of this paper. Related technologies are described in Section 2, and the proposed approach is presented in Section 3. A conclusion and discussion on future work are provided in Section 4 followed by acknowledgement and references

II. RELATED TECHNOLOGY

A. Summarisation Technology

1) Text summarisation: This refers to an automated process that produces a summary from the original content using computational techniques. The original can be a single document or multiple documents, whereas the result can be in the form of a short passage or a list of main sentences from the original document.

There are various techniques that have been applied in text summarisation. This includes statistical approach, machine learning and natural language processing (NLP). For instance, statistical approach is used to find the frequency of words or phrases from the original text in order to identify the main key words or phrases and put them as a part of the result. Supervised learning techniques can be used to identify the main key words or phrases using classifier in the training process. Such classifier may be used only in the same domain as the training data set. In addition, unsupervised learning techniques can be applied to find the main key words or phrases that has similar characteristics. Both statistical and machine learning approaches can also be applied to identify main sentences as a part of the results.

On the other hand, natural language processing techniques are other ways to produce a summary by understanding the context of the original content. Some words in the results may not come from the original content directly as they could have been referred to groups of related words provided from other sources [4]. In addition, these techniques can be mixed or integrated to produce more useful summaries such as combining natural language processing and statistical approach [5] [6].

Text summarisation can be evaluated by a comparison between the summary and the original content. On one hand, evaluation may use human to read and compare the passages. This however requires extensive time and it also relies on the competency of the human. On the other hand, text summarisation can be evaluated by using precision and recall, which are well known measurable quantities based on statistical approach in the Information Retrieval (IR) discipline. Precision refers to the measure of the correctness of the output, which refers to the relevance of the retrieved information. Recall measures the completeness of the output,
which refers to the relevant extracted information that has been retrieved.

2) **Web summarisation**: Web page is an electronic document on the World Wide Web, which is a service on the Internet. A Web page may contain various types of content such as text, picture, sound, video, interactive multimedia, and hyperlinks. These contents can be displayed in a Web browser using Hypertext Markup Language (HTML) tags, which is a standard for formatting the information in the Web page. Web pages normally provide static and/or dynamic content [7]. Static Web content refers to information on the page that is blended with HTML tag in the file and it can only be edited or modified by someone who has access to the file. On the other hand, dynamic Web contents are the information that is generated by some form of Web program using languages such as PHP and ASPX. Such information can be changed based on requests by the end users without the need to access the original file on the server. Hence this approach provides the flexibility and the ability to display information only on demand.

In order to create a summary from the Web, it is necessary to separate HTML tags and Web programming from text content and to rearrange the content into a paragraph [8]. In addition, advertisements and unrelated information have to be removed or filtered out. After the content has been extracted, text summarisation techniques can be applied. Information in meta HTML tags, which are not be displayed on the Web, can be considered as a form of summary of each Web page. Moreover, automatic indexing of outputs from services such as search engines, can also be used for indexing. Other annotations are also useful for Web summarisation as they are capable of describing the form of media on the page [9].

Web summarisation may apply the same evaluation techniques similar to text summarisation. However, there have been some other studies which compare their outputs with other summaries that have been constructed by human from DMOZ Open Directory Project [10]. In this study, the text summarisation technique is applied.

### B. Visualisation Technology

Information Visualisation is a means to explore and derive new insights on large amounts of data by visualizing the information using specific applications [11], [12]. The objective is to make the data to be easily understood by the viewers. The aims of Information visualisation are to both optimise and enhance information retrieval and presentation of huge data sets [12]. In a way, this could also be considered as a form of summary since the display may provide abstract and conceptual information from the original document.

In general, the types of visualised media should be suitably selected for knowledge transfer and they include sketches, diagrams, images, maps, objects, interactive visualisation and stories. There are a few steps on visualisation as follows.

1) - **Data preparation**: This step is for the collection of data and prepare them for the purpose of information visualisation.

2) - **Data transformation**: The collected data will be transformed and encoded to appropriate data structure.

3) - **Data visualisation**: The structured data will be applied with suitable algorithms and presented or displayed on the selected media.

Summaries can be represented visually in a group of key words, or Tag cloud [13], [14]. Tags refer to words or terms, which are extracted from the original text and they are used to represent the characteristics of the original document in the form of an image [14].

In a tag cloud, there could be different sizes, colours and styles of the importance of tags used to to draw the user’s attention with more prominent presence of the perceived important tags. The tag extraction process normally is based on statistical approaches, and the rate of recurrence is the generally used parameter for extracting the number of tags.

Visualising tag cloud could use various algorithms to layout the tags such as limited space or overlap between the tags. The tags that have relevant meaning can be clustered in groups and displayed in close-by areas [15]. Spatial clustering algorithm can be also be applied to position the tags in specific locations or particular areas [15], [16]. Some other visualization techniques also present the tag cloud in a form of circular layout [17].

### III. THE WEB ASSESSMENT APPROACH

This paper proposes a web assessment approach based on summarisation and visualisation to provide a means for users to access information from the Internet. The proposal is based on quantitative and qualitative assessments. The quantitative assessment is based on the number of objects in the webpage. This includes information such as number of words, links, images, or multimedia objects on either each Web page or the whole site. The quantity assessment is based on the context of the original text. In this paper, an integration of the existing summarisation and visualisation techniques are adopted.

The proposed assessment aims to provide different aspects of summaries of a Web page using tag cloud, text summarization and Document Type Views (DTV). This framework will allow a user to navigate the Web page and it aims to help the user to browse the content on the page and improve the quality of the returned information. The framework also provides cross-checking between different techniques. It means that it is possible to apply the key factors of one technique to be adapted into another technique. An example is the relationship between Tag cloud and Text summarization. Furthermore, the framework can also be used as a tool for Web page evaluation and comparison. It not only provides the summary and visualization of the page content, the proposed application can also present the structure of the Web pages and provides the most appropriate information according to users needed. Features of techniques used in this approach are described as follows:

#### A. Text Summarisation

This feature aims to construct a summary from the content of the Web page after the content is extracted by crawling and
pre-processing. The objective is to eliminate noise and rearrange from the original to normal text formal. In order to illustrate this component of the framework, the Extractorlive Website\textsuperscript{1} is used as an example for the extraction of the text summary used in this paper.

B. Tag Cloud

Tag cloud is another key feature of this framework for visualising abstract information from the Web’s content. This framework addresses the issue that tag cloud does not provide any meaning [13] because it provides related tags to the summary. An example of tag cloud illustrated in this paper is based on the TagCrowd Website\textsuperscript{2}.

C. Document Type View (DTV)

Document Type View is another form of visualisation. A similar idea has been demonstrated as a tool on the Ainibot Website\textsuperscript{3}. DTV provides a representative overview structure of a Web page in the form of a tree structure.

The proposed framework developed in this study also provides interactive functions. The first one is a thumbnail. When a user points to a particular hyperlink node on the DTV, the system will capture and display the information or document of that node and presents it as the thumbnail of the targeted Web page. The other function is “Drill-Down”. This can be activated by click on the hyperlink node, and the system will then retrieve the destination content and redraw the views of targeted Web page at the same time. These features will be synchronised, and an example of this framework is shown in Fig 1 below.

![Fig. 1. An example of this approach’s result](image)

IV. CONCLUSION

Due to the rapid growth of information on the Internet, information overload has caused for users more time to search for their required information. This paper proposes a qualitative and quantitative assessment approach to provide an overview of the information on a Website. The proposal is based on an integration of existing summarization and visualization techniques. These techniques are synchronised among text summarization, tag cloud, Document Type View and thumbnail as well as interactive features such as drill down. This approach is believed to reduce the time to identify the required information. In addition, the proposal can be extended as a tool for assessing and comparing Web sites.

ACKNOWLEDGMENT

The authors acknowledge the support offered by Dr. Ong Sing Goh on the Document Type View part.

REFERENCES


---

\textsuperscript{1} http://www.extractorlive.com

\textsuperscript{2} http://www.tagcrowd.com

\textsuperscript{3} http://www.ainibot.org/webtrustmetric