2010 Third International Conference on Knowledge Discovery and Data Mining

Discover Information and Knowledge from Websites using an Integrated Summarization and Visualization Framework

Chun Che Fung
School of Information Technology
Murdoch University
Western Australia, Australia
l.fung@murdoch.edu.au

Wigrai Thandechteemapat
School of Information Technology
Murdoch University
Western Australia, Australia
wigrai@ieee.org

Abstract— **The number of Web sites has noticeably increased to roughly 225 million in the last ten years. This means there is a rapid growth of knowledge and information on the Internet. Although search engines can help users to filter their desired information based on key words, the searched result is normally presented in the form of a list, and users have to visit each Web page in order to determine the appropriateness of the result. A considerable amount of time therefore has to be spent on finding the required information. To address this issue, this paper proposes a knowledge discovery approach on the Web by providing an overview of the information on a Website using an integration of summarization and visualization techniques. This includes text summarization, tag cloud, Document Type View, and interactive features such as drill down and thumbnails. This approach is capable to reduce the time required to identify and search for information or knowledge from the Web.**

Keywords-component; Text summarization, Tag Cloud, Visualization, Web assessment

I. Introduction

Knowledge and information on the Internet are expanding at a rapid growth rate and there is no central entity responsible for organising such information repository [1]. The number of Web sites has dramatically increased to approximately ten times at 225 million from 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 summarization and visualization 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 II, and the proposed approach is presented in Section III. A conclusion and discussion on future work are provided in Section IV followed by acknowledgement and references.

II. Related Technology

A. *Summarization Technology*

1) Text summarization: is 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. In addition, the result can be established by extraction or abstraction from the original text [4]. A summary by extraction is to find a part of text that can be considered as an indicative passage of the content by choosing the best ranked sentences [3]. On the other hand, a summary by abstraction aims to produce a result based on the original semantics.

There are various techniques that have been applied in text summarization. This includes statistical approach, machine learning and natural language processing (NLP). For instance, in the case of producing a summary by extraction, statistical approach can be 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 have 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 an abstracted 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 [5]. In addition, these techniques can be mixed or integrated to produce more useful summaries such as combining natural language processing and statistical approach [6] [7].

978-0-7695-3923-2/10 $26.00 © 2010 IEEE
DOI 10.1109/WKDD.2010.109

Text summarization 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 summarization 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 summarization*: 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 [8]. 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 [9]. In addition, advertisements and unrelated information have to be removed or filtered out. After the content has been extracted, text summarization techniques can be applied. Information in Meta HTML tags, which are not 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 summarization as they are capable of describing the form of media on the page [10].

Web summarization may apply the same evaluation techniques similar to text summarization. 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 [11]. In this study, the text summarization technique is applied.

B. *Visualization Technology*

Information Visualization is a means to explore and derive new insights on large amounts of data by visualizing the information using specific applications [12], [13]. The objective is to make the data to be easily understood by the viewers. The aims of Information visualization are to both optimise and enhance information retrieval and presentation of huge data sets [13]. 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 visualization and stories. There are a few steps on visualization as follows.

1) - Data preparation: This step is for the collection of data and preparing them for the purpose of information visualization.

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

3) - Data visualization: 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* [14], [15]. 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 [15].

In a tag cloud, there could be different sizes, colours and styles of the importance of tags used 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.

Visualizing 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 [16]. Spatial clustering algorithm can be also be applied to position the tags in specific locations or particular areas [16], [17]. Some other visualization techniques also present the tag cloud in a form of circular layout [18].

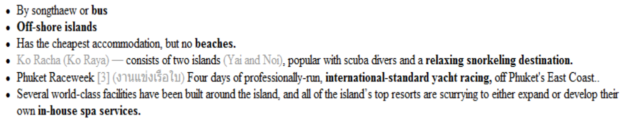
III. THE WEB ASSESSMENT APPROACH

This paper proposes a web assessment approach based on summarization and visualization 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 summarization and visualization 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 Summarization*

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, a Web page that provides information about Phuket on Wikitravel[1] is employed as example data in this framework, while the Extractorlive Website[2] is used for the extraction of the text summary used in this paper. The example result is shown in Figure 1.



- By songthaew or **bus**
- **Off-shore islands**
- Has the cheapest accommodation, but no **beaches.**
- Ko Racha (Ko Raya) — consists of two islands (Yai and Noi), popular with scuba divers and a **relaxing snorkeling destination.**
- Phuket Raceweek [3] (งานแข่งเรือใบ) Four days of professionally-run, **international-standard yacht racing,** off Phuket's East Coast..
- Several world-class facilities have been built around the island, and all of the island's top resorts are scurrying to either expand or develop their own **in-house spa services.**

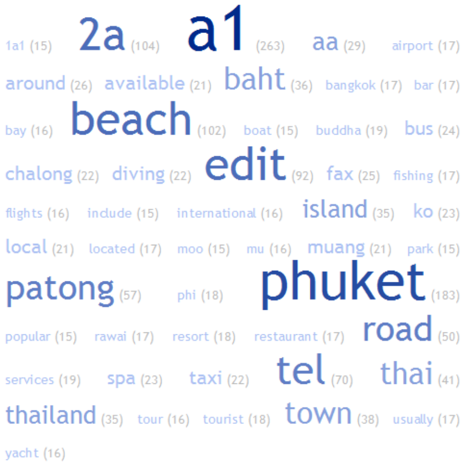
Figure 1. An example of Text summarization on Wikitravel-Phuket

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 [14] because it provides related tags to the summary. An example of tag cloud in this paper is based on the TagCrowd Website[3], and the result is shown in Figure 2.

C. *Document Type View (DTV)*

Document Type View is another form of visualization. A similar idea has been demonstrated as a tool on the Ainibot Website[4]. DTV provides a representative overview structure of a Web page in the form of a tree structure as shown in Figure 3.



1a1 (15) 2a (104) a1 (263) aa (29) airport (17)
around (26) available (21) baht (36) bangkok (17) bar (17)
bay (16) beach (102) boat (15) buddha (19) bus (24)
chalong (22) diving (22) edit (92) fax (25) fishing (17)
flights (16) include (15) international (16) island (35) ko (23)
local (21) located (17) moo (15) mu (16) muang (21) park (15)
patong (57) phi (18) phuket (183)
popular (15) rawai (17) resort (18) restaurant (17) road (50)
services (19) spa (23) taxi (22) tel (70) thai (41)
thailand (35) tour (16) tourist (18) town (38) usually (17)
yacht (16)

Figure 2. An example of tag cloud on Wikitravel-Phuket

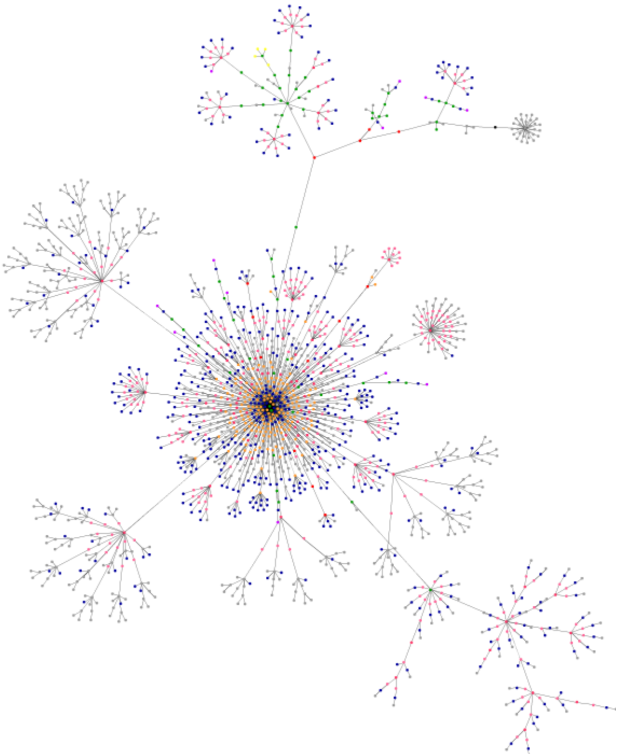
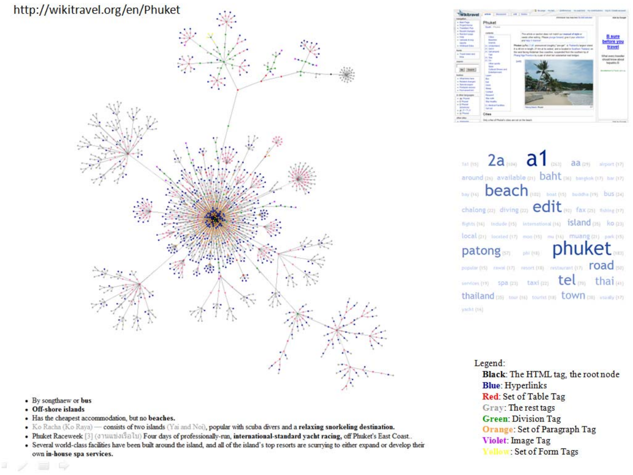


Figure 3. An example of Document Type View on Wikitravel-Phuket

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 synchronized, and an example of this framework is illustrated in Figure 4.

[1] http://wikitravel.org/en/Phuket
[2] http://www.extractorlive.com
[3] http://www.tagcrowd.com
[4] http://www.ainibot.org/webtrustmetric


http://wikitravel.org/en/Phuket
1a1 (15) 2a (104) a1 (263) aa (29) airport (17)
around (26) available (21) baht (36) bangkok (17) bar (17)
bay (16) beach (102) boat (15) buddha (19) bus (24)
chalong (22) diving (22) edit (92) fax (25) fishing (17)
flights (16) include (15) international (16) island (35) ko (23)
local (21) located (17) moo (15) mu (16) muang (21) park (15)
patong (57) phi (18) phuket (183)
popular (15) rawai (17) resort (18) restaurant (17) road (50)
services (19) spa (23) taxi (22) tel (70) thai (41)
thailand (35) tour (16) tourist (18) town (38) usually (17)
yacht (16)
Legend:
Black: The HTML tag, the root node
Blue: Hyperlinks
Red: Set of Table Tag
Gray: The rest tags
Green: Division Tag
Orange: Set of Paragraph Tag
Violet: Image Tag
Yellow: Set of Form Tags

- By songthaew or **bus**
- **Off-shore islands**
- Has the cheapest accommodation, but no **beaches.**
- Ko Racha (Ko Raya) — consists of two islands (Yai and Noi), popular with scuba divers and a **relaxing snorkeling destination.**
- Phuket Raceweek [3] (งานแข่งเรือใบ) Four days of professionally-run, **international-standard yacht racing**, off Phuket's East Coast..
- Several world-class facilities have been built around the island, and all of the island's top resorts are scurrying to either expand or develop their own **in-house spa services.**

Figure 4. An example result of this approach on Wikitravel-Phuket

IV. CONCLUSION

Due to the rapid growth of knowledge and information on the Internet, information overload may cause the users to use more time to search for their required information. This paper proposes a qualitative and quantitative 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 synchronized among text summarization, tag cloud, Document Type View and thumbnails as well as interactive features such as drill-down. This approach is believed to be able to assist in discovery of knowledge and information while reducing 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 section as regard to Document Type View.

REFERENCES

[1] Rohit, Kaul, et al., *Ranking billions of web pages using diodes*. Commun. ACM, 2009. **52**(8): p. 132-136.

[2] Netcraft. *August 2009 Web Server Survey*. 2009 31 August 2009 [cited 2009 Sep 6]; Available from: http://news.netcraft.com/archives/2009/08/index.html.

[3] Huantong, Geng, et al. *A Novel Automatic Text Summarization Study Based on Term Co-Occurrence*. in *Cognitive Informatics, 2006. ICCI 2006. 5th IEEE International Conference on*. 2006.

[4] Sornil, O. and K. Gree-ut. *An Automatic Text Summarization Approach using Content-Based and Graph-Based Characteristics*. in *Cybernetics and Intelligent Systems, 2006 IEEE Conference on*. 2006.

[5] Jiang-Liang, Hou and A. W. J. Tsai, *Knowledge Reuse Enhancement with Motional Visual Representation.* Knowledge and Data Engineering, IEEE Transactions on, 2008. **20**(10): p. 1424-1439.

[6] Zhang, Y. Z., et al. *Summarizing Web Sites Automatically*. in *In Proc. Conference of Canadian Society for Computational Studies of Intelligence*. 2003.

[7] Amitay, Einat and Paris Cecile, *Automatically summarising Web sites: is there a way around it?*, in *Proceedings of the ninth international conference on Information and knowledge management*. 2000, ACM: McLean, Virginia, United States.

[8] Thanadechteemapat, Wigrai and Chun Che Fung. *A Survey on the Use of Web Technologies in the Promotion of Sustainable Energy* in *the 9th Postgraduate Electrical Engineering & Computing Symposium (PEECS 2008)*. 2008. Perth.

[9] Fung, Chun Che, et al. *iWISE, an intelligent Web Interactive Summarization Engine*. in *Machine Learning and Cybernetics, 2009 International Conference on*. 2009.

[10] Kobayashi, Mei and Takeda Koichi, *Information retrieval on the web.* ACM Comput. Surv., 2000. **32**(2): p. 144-173.

[11] *About the Open Directory Project*. 2002 [cited 2009 27 Feb]; Available from: http://www.dmoz.org/about.html.

[12] Burkhard, Remo Aslak, *Knowledge Visualization - The Use of Complementary Visual Representations for the Transfer of Knowledge. A Model, a Framework, and Four New Approaches*. 2005, Swiss Federal Institute of Technology Zurich.

[13] Eppler, Martin J and Remo A. Burkhard, *Knowledge Visualization*. 2004, NetAcademy Project: Switzerland.

[14] Hearst, M. A. and D. Rosner. *Tag Clouds: Data Analysis Tool or Social Signaller?* in *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*. 2008.

[15] McKie, S. *Scriptclud.com: Content Clouds for Screenplays*. in *Semantic Media Adaptation and Personalization, Second International Workshop on*. 2007.

[16] Rivadeneira, A. W., et al., *Getting our head in the clouds: toward evaluation studies of tagclouds*, in *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2007, ACM: San Jose, California, USA.

[17] Slingsby, A., et al. *Interactive Tag Maps and Tag Clouds for the Multiscale Exploration of Large Spatio-temporal Datasets*. in *Information Visualization, 2007. IV '07. 11th International Conference*. 2007.

[18] Seifert, C., et al. *On the Beauty and Usability of Tag Clouds*. in *Information Visualisation, 2008. IV '08. 12th International Conference*. 2008.