Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009

IWISE, AN INTELLIGENT WEB INTERACTIVE SUMMARIZATION ENGINE

CHUN CHE FUNG[1], WIGRAI THANADECHTEEMAPAT[2], KIT PO WONG[3]

[1, 2] School of Information Technology, Murdoch University, Murdoch 6150, Western Australia
[2]Department of Electrical Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
E-MAIL: l.fung@murdoch.edu.au , wigrai@ieee.org, eekpwong@polyu.edu.hk

Abstract:

The Internet is expanding at a rapid rate and in particular the World Wide Web (WWW) has a huge number of users and providers of information. There are many web pages of information being added and the task of finding the most appropriate information is getting difficult. While search engines are capable to return a collection of links according to key terms and some form of ranking mechanism, the users are still need to access the Web page and navigate through the site in order to find the information. This paper proposes an interactive summarization framework called iWISE to facilitate the process. The details and background the research work are provided with discussion on future work.

Keywords:

Web Summarization; Tag Cloud; Document Type View; iWISE; Visualization; Summarization Engine

1. Introduction

The Internet is expanding at a rapid rate. One of the services on the Internet which has a high growth rate is the World Wide Web (WWW). WWW acts as a convenient means of communication by serving electronic documents called Web pages through "Web sites". These sites are effectively collections of Web pages together serving the documents either on the same page, or through "links" to other pages. The number of Web sites has dramatically increased to approximately 224 million, or around ten times between 1996 and 2009 [1]. This indicates that there is an enormous amount of Web pages and document available online. This has led to the phenomenon of "*information overload*". The consequences of this can be two-fold. First, one may not know the existence of the information or the appropriate Web pages. Secondly, even when directed to the correct Web site, one still needs time to find out the exact location of the piece of information. Obviously, the implies that much time has been wasted in searching for information and locating the right information.

Even though there are many search engines addressing the first issue and are able to provide search results in a few seconds, the user still faces with a large amount of "links" without knowing the suitability of the information unless the sites are being visited one by one. What it means is, even the search results are based on the key words being entered, the user still has to access the information from the Web and incur the associated traffic costs. As for the second issue, the problem is due to the fact that the search engines do not return the results based on the content or information in the sites. The user has no way to know the appropriateness and the ranking will depend on the algorithm deployed by the particular search engine. For these reasons, it is necessary to derive new ways to access the information and provide a more comprehensive overview to the users. Web summarization is the approach being proposed and discussed in this paper. It is believed that the proposal will address the second issue and will reduce the time required by the user to look for the appropriate information in a Web site.

There have been many reports on research work related to various summarization techniques. The objectives of such techniques are mainly to produce a coherent summary as concise as it could be. However, it is believed that a summary should not only present a coherent summary, but it should also be interactive and be able to provide different aspects of the information. As a result, this paper proposes an alternative summarization approach, which integrates existing summarization and visualization techniques in order to deliver different aspects of a Web site. The interactively generated Web summary should be flexible and allows navigation between the different aspects of the information and to improve the appropriateness of the information being sought. The proposed approach is termed iWISE, an intelligent Web Interactive Summarization Engine.

This paper starts with an introduction on background and the aims of this paper. This is followed by a brief description on summarization technology being deployed in iWISE. They are followed by other related work and a detailed description of the features of iWISE. Finally, this paper concludes with discussions and direction for future

978-1-4244-3703-0/09/$25.00 ©2009 IEEE

work.

2. Summarization Technology

There are many methodologies and techniques on summarization. In this section, only related technologies to iWISE will be explained.

2.1 Text Summarization

Text summarization is based on a variety of techniques. The most common one is the statistical approach such as frequency and histogram of key words. For example, frequency is used in the first part of the summarization process for indentifying frequent occurrence of words from the original source. The words are then combined in a summary.

Natural language processing is the other technique being applied to text summarization. It can work in a pre-processing stage for extracting key phrases and indentify them with respect to related domain [2]. Some tasks on natural language processing have similar characteristics of Information extraction (IE) [3]. They are used for finding the most appropriate information [4] to form a part of summaries. Moreover, natural language processing can be combined with statistical approaches to form one paragraph or text section from a passage [5].

2.2. Tag Cloud

A "tag" is related to words or terms extracted from an original passage or data, whereas "tag cloud" is a group of tags that represents the characteristics of the original data in a form of an image [6]. This is a form of visualization or summary [7, 8] as they provide abstract information from the original text data. In a tag cloud, the importance of the tag could be illustrated or visualized according to size, color and style. The idea is to draw the attention of the user with more prominent display of the perceived important tags. The process of tag extraction normally is based on statistical approaches. Frequency is the commonly used parameter in the early step for gathering the number of words or terms.

Visualization of tag cloud could be based on different algorithms to layout the tags such as fixing the amount of white space or overlap between the tags. Clustering is also used to group the tags that have related meanings in the closed by area [9]. On the other hand, spatial clustering algorithm can be applied to arrange the tags in particular areas or locations [9, 10]. Some visualization techniques aim to present tag cloud in other ways such as inclusion of their relationships, or in the form of circular layout [7].

2.3. Summarization on Web pages

Web page is a form of electronic document. A Web page's content can be text with color and font attributes, picture of various format, voice, sound, video and hyperlinks. In addition, a Web page may provide interactivity with the users. Traditionally, most of the Web pages are formatted using Hypertext Markup Language (HTML) and displayed through Web browser. That means that the information on a Web page is blended with markups or tags in HTML [11]. Some content on Web page, which is called metadata, are not presented to users. Such metadata is used to describe the information on the Web page using metatags in HTML. Moreover, metadata can be collected in a process for automatic indexing [12]. As a result, metadata can be a form of summary on each Web page.

Annotation is another part of a Web page's content which is used to provide indexes. Web page content evaluation based on annotation may be useful because it contains additional information. Annotation can therefore be used in the automatic summarization process and for content-based retrieval since annotation may describe the multimedia content on the Web page[12].

Summarizing the information on Web page has more processes than text summarization. Not only HTML tags have to be removed, but the information without HTML tags must be rearranged before commencing the process of summarization. These steps have to be preceded subsequent steps for Web mining. In this paper, Web mining however is not covered as this forms another part of the project.

3. Other Work on Summarization

Berger and Mittal [13] proposed a prototype system called OCELOT, which can summarizes content on the Web without selecting either sentences or paragraphs. This is known as an extractive summarization technique. OCELOT is based on the Open Directory Project [14]. The Open Directory Project is a Web site providing multi language directory with short description edited by volunteers. Berger and Mittal used the project Web site as the data source and a part of the evaluation process.

The OCELOT's technique is related to statistical machine translation based on natural language and machine learning. There are three steps in this system. The first step determines the words that should be in the summary. This is called *content selection* and it is based on statistical

techniques. *Word ordering* is the second step, which arranges the words in the summary. The last step is *search*, which uses the arranged words and apply a search technique, the Viterbi algorithm [15], to search the arranged words and then form the summary. Berger and Mittal described some of the limitations of the system. For example, it cannot distinguish between subject or object in sentences. The phrases "Dog bites man" and "Man bites dog" will be treated the same although it is obvious the meanings are different. In summary, the system provides text summarization from Web's content, but Web users may need to spend some time to read and comprehend the summary.

Baratis and et al. [16] presented an automated image-based summarization approach, which can be considered as a Web site summarization based on image content. They use logo and trademark images as the basis of the summarization. This approach is useful to detect unauthorized uses of trademark and logo, so it has commercial significance and has received much attention. There are four steps in this approach. *Image information extraction* is the first step, which extracts related information based on the images. Next, *machine learning*, based on decision trees, is used to distinguish the extracted images. Such images are also used for training the system after the images were converted to gray scale. This step is called the *Logo and trademark detection*. The images are then processed and detected for duplication of logo and trademark. These images are used to create *clusters*. This is the third step. Each image is ranked according to the frequency of occurrence, number of hyperlinks and other information. The last step is *image-based summarization*, which ranks each cluster, and the high rank clusters will be displayed in the summary. However, this approach would have been better if it can provide short descriptions and is able to interact with the users by allowing them to offer a concise summary on the images. In addition, the system should provide real time service.

Amitary and Paris [5] introduced InCommonSense, which summarizes Web pages in an alternative approach that relies on hypertext structure and its information provided by the Web authors or Webmasters. The technique is considered by Amitary and Paris as a "*fully automatic pseudo-summarization technique"*.

The system starts by taking one Web page and then finds other Web pages that are linked back to the original Web page. Each of the other Web pages will be examined and information on any anchor that is linked to the required page will be extracted. Only one of the anchors and their information will be selected. The selection and rating of each anchor description is based on the machine learning tool, See5 [17], with 16 hard coded rules. The system presents results in a form of merging short summary to results from search engines.

Zhang and et al. [18] proposed another approach to summarize Web site automatically. Natural language processing and machine learning are applied in this approach. This approach has five steps to produce the summary. The first step will start from home page. The next step is to extract all the contents from each Web page into paragraph format. The third step is to filter out *short* paragraphs and produce classifiers from the rest of the paragraphs into *narrative* or *non-narrative* groups extracted by shallow natural language processing and machine learning. The fourth step extracts the key-phrases by the trained classifiers into categories of phrases. The last step extracts the most important sentences based on the occurrence of key-phrases, from the narrative paragraphs to assemble a final summary. In addition, the criteria proposed by Zhang and et al on the generated summary, based on this approach, should comprises of the top 25 key-words, top 10 key-terms and top 5 key-sentences. The next section provides a description of the iWISE framework.

4. Intelligent Web Interactive Summarization Engine (iWISE)

A novel approach called iWISE is proposed. Its key function is to provide on Web summarization and it is based on an integration of existing summarization and visualization techniques. Visualization is used to represent abstract information or concept in an image or a graph [19]. iWISE aims to provide different aspects of summaries of a Web page using tag cloud, text summarization and Document Type Views (DTV). The proposed system is interactive and may operate in real time. Also, iWISE allows users to navigate the Web page easily and to reduce time in browsing information on the Web thereby improving the quality of the returned information. It also permits cross-checking between different techniques. That means that key factors of one technique could be adapted into another technique. An example is the connection between Tag cloud and Text summarization.

In addition, iWISE can be used as a tool for evaluating Web pages since iWISE not only provides the summary and visualization of the Web's content, but it can also also present the structure of the Web pages. Some features of iWISE are described as follows:

Text Summarization

This feature aims to produce a summary from the Web page, and the summary may be in the form of highlighted sentence lists or summary paragraphs. In this paper,

www.extractorlive.com is used as an example for illustration. Result from the web site, www.pcworld.com, is shown in Figure 1.

Tag Cloud

iWISE uses tag cloud as a part of the key output from the summarization engine, and an example of tag cloud in this paper is obtained from the website TagCrowd. While tag cloud is recognized that it does not provide any meaning [8], iWISE provides the relevant results from text summarization to address this issue.

Figure 1. Home page of www.pcworld.com

Document Type View (DTV)

iWISE represents Web page in a form of visualization, which is similar to Webtrustmetrix being used on Ainibot Web site. Document Type View can depict an overview of the nature of the information on a Web page. An exam,ple of DTV based on the PC World Web site is shown in Figure 2.

Furthermore, DTV provides an interactive feature known as "Drill-Down". By clicking the hyperlink node, iWISE will download and extract the destination Web page and redraw the view of the targeted Web page. This feature allows a user to go to any link on the Web page and find out the DTV of that node.

iWISE provides an integrated view based on text summarization, tag cloud and Document Type View at the same time. The features will be synchronized, and the result of the site www.pcworld.com shown in Figure 3.

As for the Drill-Down function, iWISE will repeat the same working process while text summarization and tag cloud are generated at the same time as shown in Figure 4.

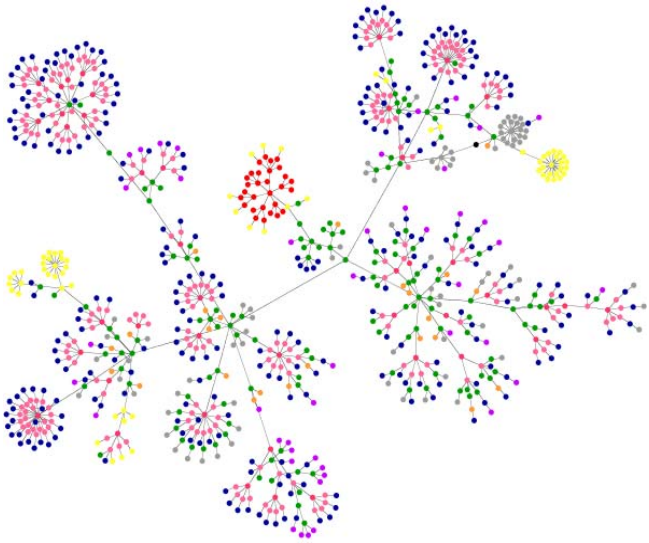

Figure 2. Document type view of iWISE

5. Discussion and Conclusion

There are a large number of Web pages that provides information on the Internet. This will require more and more time to search for the required information on the Internet. Therefore, it is useful to have a new way that provides a more comprehensive view of the Web content to the users. This paper has proposed an intelligent Web Interactive Summarization Engine (iWISE), which is a novel approach on Web Summarization based on an integration of existing summarization and visualization techniques. The implementation and experimentation are in progress and further results will be reported in due course.

Future development on iWISE will include machine learning in order to serve the individual users with the characteristics of customization and personalization. Also, iWISE may also be used for ranking system based on quantitative assessment of the web contents. Other work on iWISE will include objective assessments based on performance, user perspective, usability and acceptance.

Acknowledgements

The authors acknowledge the assistance offered by Dr. Ong Sing Goh on the section as regarded to Document Type View.

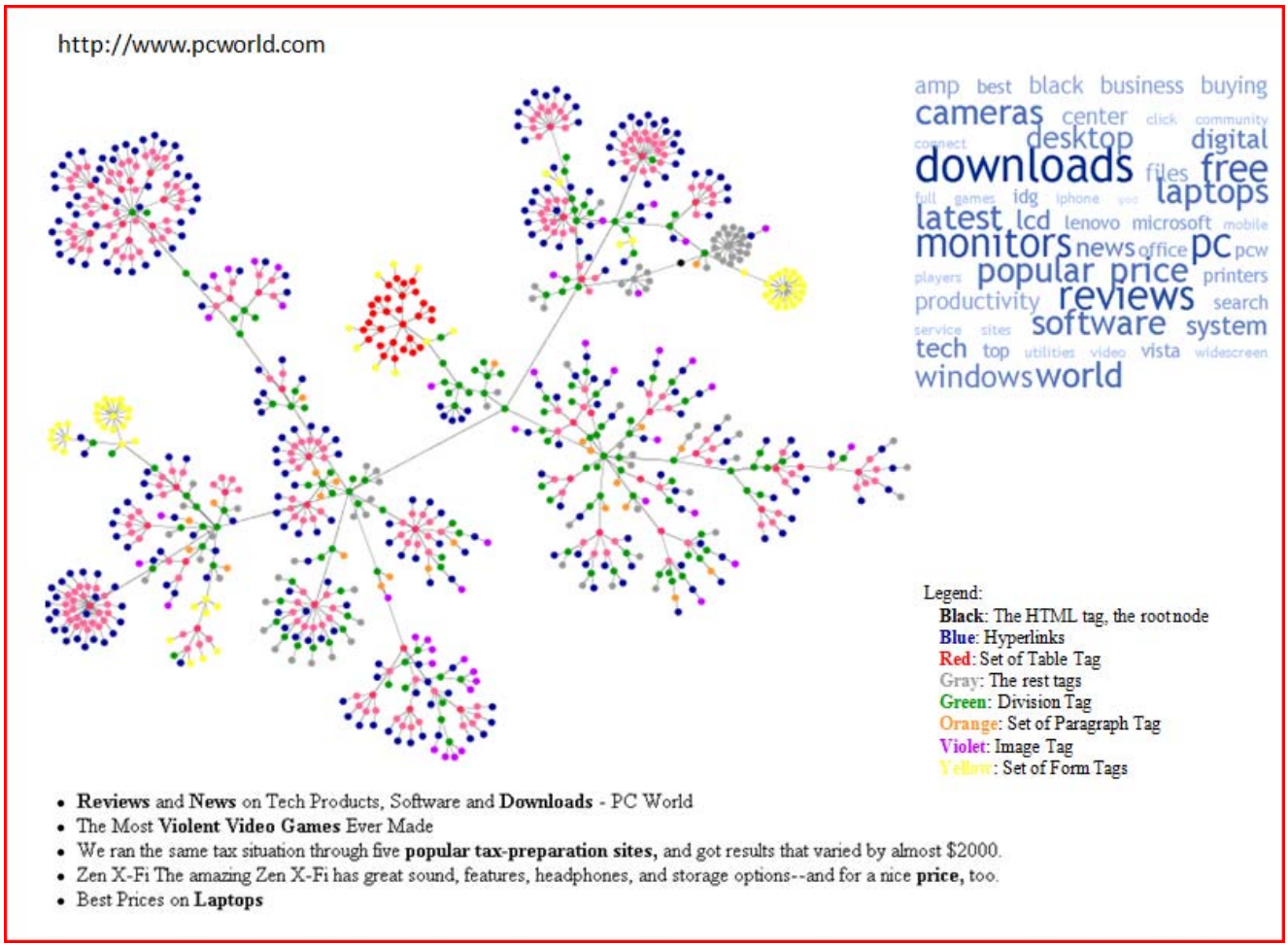

Figure 3. User View of iWISE

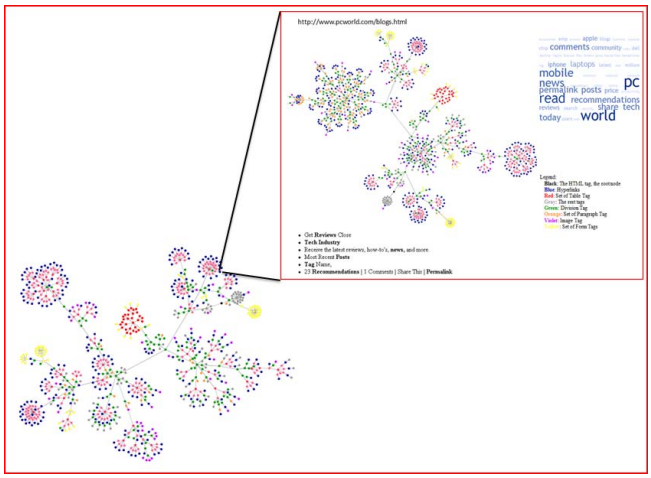

Figure 4. Illustrating Drill-Down feature of iWISE

References

[1] Netcraft. *March 2009 Web Server Survey*. 2009 2 March 2009 [cited 2009 Mar 28]; Available from: http://news.netcraft.com/archives/2009/03/index.html.

[2] 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.

[3] Spärck Jones, Karen, *Automatic summarising: The state of the art.* Information Processing & Management, 2007. 43(6): p. 1449-1481.

[4] Jim, Cowie and Lehnert Wendy, *Information extraction.* Commun. ACM, 1996. 39(1): p. 80-91.

[5] 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.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] Berger, Adam, L. and O. Mittal Vibhu, *OCELOT: a system for summarizing Web pages*, in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. 2000, ACM: Athens, Greece.

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

[15] Forney, G. D., Jr., *The viterbi algorithm.* Proceedings of the IEEE, 1973. 61(3): p. 268-278.

[16] Baratis, Evdoxios, et al., *Automatic Website Summarization by Image Content: A Case Study with Logo and Trademark Images.* Knowledge and Data Engineering, IEEE Transactions on, 2008. 20(9): p. 1195-1204.

[17] Quinlan, J. Ross, *C4.5: programs for machine learning*. 1993: Morgan Kaufmann Publishers Inc. 302.

[18] Zhang, Y., *World Wide Web site summarization.* Web Intelligence and Agent Systems, 2004. **2**(1): p. 39-53.

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