http://dx.doi.org/10.1049/cp.2009.1852

SUMMARIZING INFORMATION FROM WEB SITES ON DISTRIBUTED POWER GENERATION AND ALTERNATIVE ENERGY DEVELOPMENT

Chun Che Fung

School of Information
Technology,
Murdoch University,
WA 6150, Australia

Wigrai Thanadechteemapat

School of Information
Technology,
Murdoch University,
WA 6150, Australia

Kit Po Wong

Department of Electrical
Engineering, The Hong Kong
Polytechnic University,
Hung Hom, Hong Kong

ABSTRACT

The World Wide Web (WWW) has become a huge repository of information and knowledge, and an essential channel for information exchange. Many sites and thousands of pages of information on distributed power generation and alternate energy development are being added or modified constantly 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 forms of ranking mechanism, it is still necessary 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 by providing a summary of the information on the Web site. The proposed approach makes use of graphical visualization, tag clouds and text summarization. A number of cases are presented and compared in this paper with a discussion on future work.

1. INTRODUCTION

The world is facing an urgent challenge of how to develop alternative energy and distributed power generation efficiently. Most of the world's energy demands at present are still relying on crude oil, coal and natural gas [1]. It can be said that such resources are finite and there are limitations on the supply side of the equation and eventually such resources will not able to meet the upward demands. In 2006, it has been reported that the total world production of energy was already unable to maintain further growth in demand [2]. Also, the price of crude oil has undergone extreme fluctuations [3]. For example, the oil prices in 2008 had been rapidly increased in the first half and then dropped steeply in the second half of the year [4]. This has a significant impact on the world economy [5]. On the other hand, environmental issues due to 'greenhouse effects' as a consequence of CO_2 emission from fossil fuel consumption, may have contributed towards the global climate change [6]. That will affect many generations to come and it is predictable that this will have enormous social, economic and environmental implications [8]. Hence, there is an urgent need to develop alternate energy and distributed power generation which has to be environmental friendly, sustainable, reliable and hopefully, cheap.

One of the solutions is to reduce the world's dependence on fossil fuels by escalating the reliance on alternative or sustainable energy sources. Many governments, research and commercial organizations have already committed considerable amount of effort and resources to this cause. Such effort can be further facilitated by providing and exchanging up-to-date information and knowledge among the researchers and related bodies in the discipline. An effective means for the provision and distribution of information cheaply and quickly is to use the World-Wide-Web on the Internet. At present, the number of Web sites has significantly increased approximately ten times from 1996 to 2009 at 224 million Web sites [9]. Information on alternative energy have already been provided by many Web sites as indicated in a survey study by the authors [10]. However, the enormous amount of information on the Web leads to the phenomenon of "information overload" and this has caused much burden on the researchers [11]. An inefficient process or inability of locating the required information may lead to wasted efforts and hinders the promotion and development of alternative energy. Although search engines are currently the de-facto tool that most searches on the web are based on, a large amount of results are normally produced due to information overloading. Users still have to visit each individual Web page in order to identify whether it contains the desired information. On the other hand, search engines also deployed their own algorithms or mechanisms in ranking the results and some of them may be based on financial interest. Hence, the results do not provide any context or quality of the sites. As a solution to these problems, a new approach termed iWISE, *an intelligent Website Information Summarization Engine*, is proposed in this paper. The proposed approach presents the information from the web sites with a comprehensive overview by using visualization and summarization techniques. This approach is believed having the potential to reduce the time to identify and traverse the information in a Web site.

This paper is organized with an introduction and aims of this paper. Summarization Technology is then described in Section 2, and visualization and summarization engine are explained in Section 3. Application towards alternative energy information on the web is illustrated in Section 4, and the conclusion including discussion on future work in Section 5 followed by acknowledgement and references.

2. SUMMARIZATION TECHNOLOGY

A summary is used to provide the key meaning of an original source of information in a reduced and concise representation. This representation could be in the forms of graphics, tables or text. Computer programs using various intelligent techniques may also be used to create a summary. A number of the related methodologies and techniques in iWISE are explained in the following subsections.

2.1 Text Summarization

Text summarization is a process used to produce a summary based on the original information. The summary can be produced by extraction or abstraction from the original text [12]. A summary by extraction is to find a part of text that can be considered as indicative of the content using choosing the best rank of sentences [11]. The most common technique is the statistical approach that examines sentences, phrases, or words based on frequency or histogram. They are then ranked and the highest rank sentences and phrases are used to produce a summary. There are a diversity of such techniques being developed and applied.

The other approach is summary by abstraction, which aims to a summary using natural language processing (NLP) techniques based on the original semantics. NLP attempts to understand and process human languages by extracting key phrases of the original text [13]. It can also be applied with machine learning using both supervised and unsupervised learning approaches. For example, the original text and its key phrases can be set as a training set in the form of a classifier, which determines a threshold for choosing the key phrases. This will set limits on different domains and the classifier has to be defined with some rules in each domain with specific knowledge, and a number of training data sets are needed.

On the other hand, unsupervised learning can also be used as a part of the summarization process. One of the techniques is based on an algorithm to examine the original text in order to identify key phrases, which represent most of the main meaning in the original document.

Furthermore, there is a hybrid approach which applies both statistics and natural language processing for producing useful summaries [14].

2.2. Tag Cloud

Tag cloud is a group of tags that are extracted from related words in the original text [15]. It can represent the original information in a form of an image of short phrases or words, which is considered to be a form of visual summary. Tags can be extracted by applying statistical approaches such as finding frequency of words. Unsupervised learning techniques such as clustering can be applied to identify and select the words that are properly related to the original document [16].

The cloud normally shows the extracted tags in different illustration according to style, color and size from the underlining words or phrases in the original text. Visualization of the cloud may use diverse algorithms to fit the words in one limited layout. The limited area can be fitted the tags by using spatial clustering algorithm [16] [17]. On the other hand, the cloud may also be represented in different ways such as in a circular layout [18]. Further filtering of the tags could be applied based on specific domain and elimination of non-representative words.

2.3. Summarization of Web pages

Web pages in electronic format can provide better interactive and visual information than normal print based documents. The information on the Web page can include sound, video, hyperlinks, and interactivity. Web page uses Hypertext Markup Language (HTML) to format its content in a normal text file consists of HTML tags and content. The latest Web pages may also include the use of Extended Hypertext Markup Language (XHTML), Extensible Markup Language (XML) and Script languages. The formatted contents are required to be displayed in a browser. If the content is predominately text, it can be summarized by text summarization techniques. A summary of the information on the Web page has more steps than text summarization due to the inclusion of HTML tags and other script languages which have to be removed first. Furthermore, the content also has to be rearranged before working on the summarization process. For example, the content that is displayed in tables needs to be carefully restructured.

On the other hand, metadata in a Web page can be used as a form of its summary because the metadata is used metatags to provide a short description.

Normally, the metadata are not displayed and they can be used to facilitate an automatic indexing process [19].

3. iWISE - INTELLIGENT WEBSITE INFORMATION SUMMARIZATION ENGINE

This paper proposes a novel approach called iWISE, which integrates summarization techniques with visualization techniques in order to provide a summary of a Website. iWISE provides various aspects of summaries of a Website using tag cloud, text summarization and Document Type Views (DTV) as well as a thumbnail of the Web pages within the site. The engine operates in real time and it is interactive, whereby a user can navigate the Web page through this engine. It also allows a form of cross checking between the techniques to enable the users to gain a higher degree of confidence on a summary. For instance, most of words in tag cloud should match to the key phrases in a summary from text summarization. Moreover, iWISE can be used as a tool for comparing Web sites in order to select the Web pages which deem to provide the most appropriate information according to needs of the users.

Text Summarization is a feature aiming to produce a summary of the Web page, while the summary may be highlighted sentence lists or paragraphs. The summarization engine provided from www.extractorlive.com is used in this paper as an example of this approach. The home page of ScottishPower Renewables and the result of text summarization are shown in Figure 1 and 2, respectively.

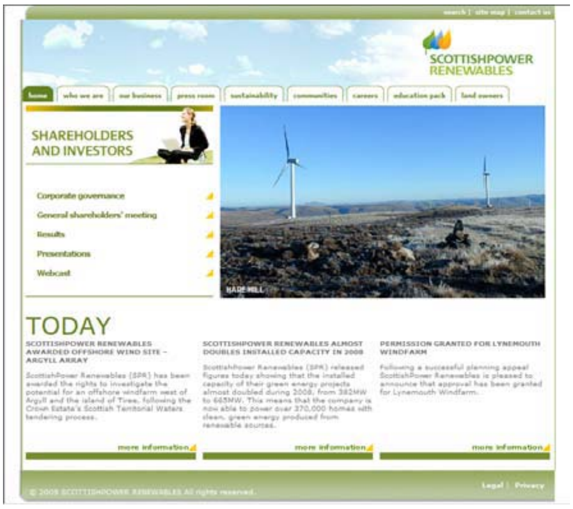
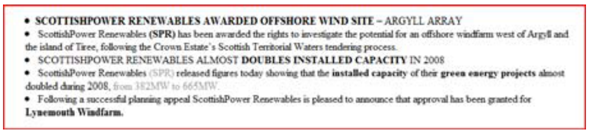

SCOTTISHPOWER RENEWABLES
SHAREHOLDERS AND INVESTORS
Corporate governance
General shareholders' meeting
Results
Presentations
Webcast
TODAY
SCOTTISHPOWER RENEWABLES AWARDED OFFSHORE WIND SITE – ARGYLL ARRAY
SCOTTISHPOWER RENEWABLES ALMOST DOUBLES INSTALLED CAPACITY IN 2008
PERMISSION GRANTED FOR LYNEMOUTH WINDFARM
more information
more information
more information

Figure 1. Home page of ScottishPower Renewables[1]



- **SCOTTISHPOWER RENEWABLES AWARDED OFFSHORE WIND SITE** – ARGYLL ARRAY
- ScottishPower Renewables **(SPR)** has been awarded the rights to investigate the potential for an offshore windfarm west of Argyll and the island of Tiree, following the Crown Estate's Scottish Territorial Waters tendering process.
- SCOTTISHPOWER RENEWABLES ALMOST **DOUBLES INSTALLED CAPACITY** IN 2008
- ScottishPower Renewables (SPR) released figures today showing that the **installed capacity** of their **green energy projects** almost doubled during 2008, from 382MW to 665MW.
- Following a successful planning appeal ScottishPower Renewables is pleased to announce that approval has been granted for **Lynemouth Windfarm.**

Figure 2. A summary of the home page of the ScottishPower Renewables

Tag Cloud is another key summarization output in iWISE. The example used in this paper is based on results from the TagCrowd website[2]. While it is recognized that there are no specific meanings on tag clouds [20], iWISE can address this issue by linking the related results to text summarization. The example tag cloud of the ScottishPower Renewables is shown in Figure 3 below

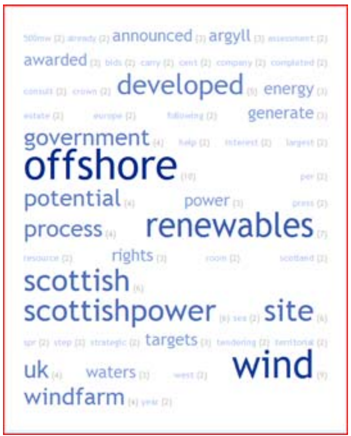


Figure 3. A tag cloud of the home page of the ScottishPower Renewables

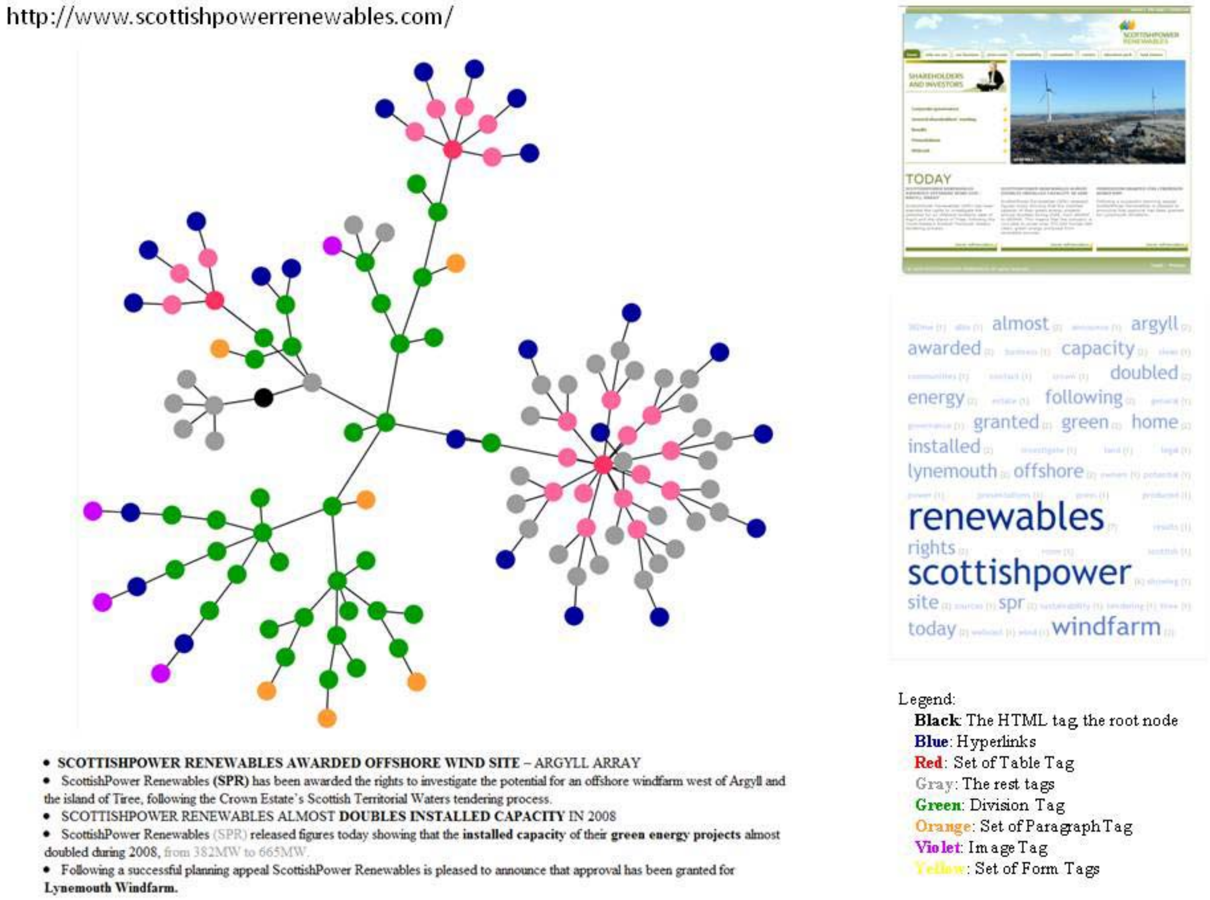
Document Type View (DTV) is a visual summary output from the iWISE approach. This is comparable to the Webtrustmetrix developed by one of the authors used on the Ainibot Web site[3]. This feature is used to provide an illustration on the general structure of a Website.

The DTV result of the ScottishPower Renewables Website is shown in Figure 4. In addition, a small thumbnail of the Web page can be displayed alongside with the overview of the Web page when a user points to the hyperlink node on the DTV. This is illustrated in Figure 5. Thus, a user may visit any Web page that is displayed in the DTV by the "Drill Down" function. When the linked node has been drilled down, the engine will repeat the working process to produce text summary, tag cloud and Document Type View, which will be synchronized, at the same time. An example of a drilled down page is shown in Figure 6.

[1] http://www.scottishpowerrenewables.com

[2] http:// www.tagcrowd.com

[3] http://www.ainibot.org/webtrustmetric



- **SCOTTISHPOWER RENEWABLES AWARDED OFFSHORE WIND SITE** – ARGYLL ARRAY
- ScottishPower Renewables **(SPR)** has been awarded the rights to investigate the potential for an offshore windfarm west of Argyll and the island of Tiree, following the Crown Estate's Scottish Territorial Waters tendering process.
- SCOTTISHPOWER RENEWABLES ALMOST **DOUBLES INSTALLED CAPACITY** IN 2008
- ScottishPower Renewables (SPR) released figures today showing that the **installed capacity** of their **green energy projects** almost doubled during 2008, from 382MW to 665MW.
- Following a successful planning appeal ScottishPower Renewables is pleased to announce that approval has been granted for **Lynemouth Windfarm.**

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

Figure 4. An actual result in this approach

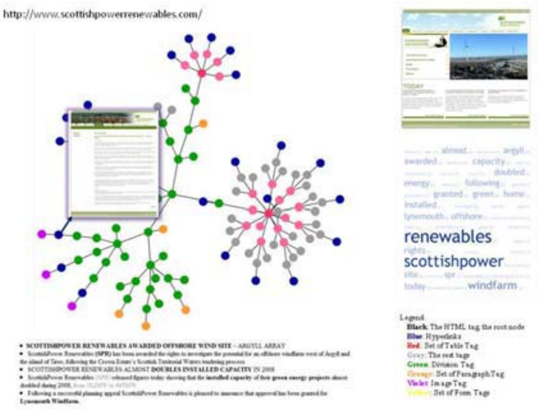


Figure 5. A thumbnail on a linked node is shown in the drill down feature

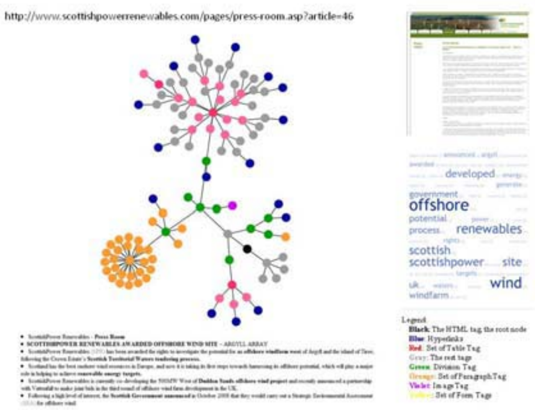

http://www.scottishpowerrenewables.com/pages/press-room.asp?article=46
developed
offshore
renewables
scottish
scottishpower
wind

Figure 6. Illustration of a result after clicking a linked node

4. ALTERNATIVE ENERGY INFORMATION APPLICATION

In this paper, four Web sites containing information on power generation and alternative energy development have been investigated.

ScottishPower Renewables[4] is a part of Iberdrola Renewables, which is one of the biggest producers of renewable energy of the world. This Web site has also been used in the earlier examples and shown in Figures 1 to 6.

China Light and Power (CLP) group[5] is one of the largest electricity investor and operators in the Asia Pacific region. The result of the engine is shown in Figure 7.

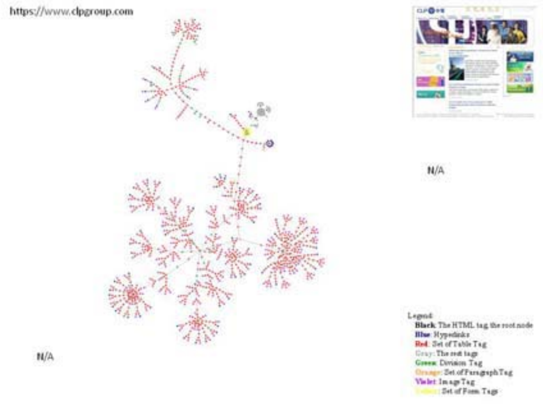
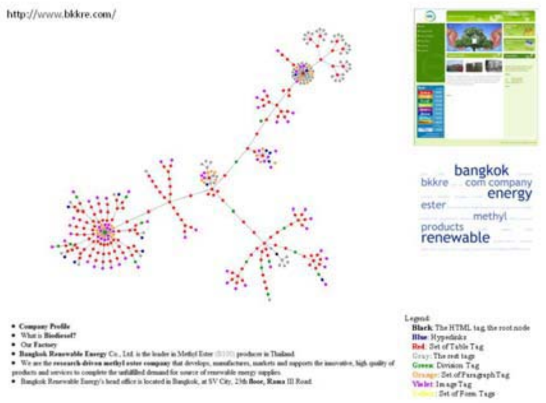


Figure 7. View of home page of CLP group

In this figure, text summarization and tag cloud could not be generated since both ExtractorLive and TagCrowd do not support Secure Sockets Layout (SSL), which has been used at the Web site of CLP group.

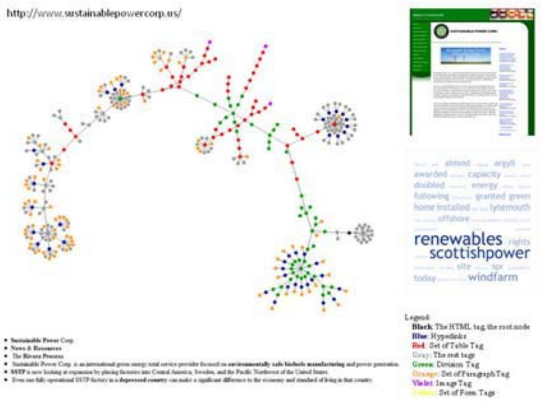
Bangkok Renewable Energy[6] is a leader producer in Methyl Ester (B100) in Thailand. The result of summarization of the site is shown in Figure 8.

Sustainable Power Corp (SSTP)[7] produces biofuel and focuses on turnkey development and management of independent bio fuel manufacturing and power plants utilizing green energy facilities. The result of summarization of this site is shown in Figure 9.



- **Company Profile**
- What is **Biodiesel?**
- Our Factory
- **Bangkok Renewable Energy** Co., Ltd. is the leader in Methyl Ester (B100) producer in Thailand.
- We are the **research driven methyl ester company** that develops, manufactures, markets and supports the innovative, high quality of products and services to complete the unfulfilled demand for source of renewable energy supplies.
- Bangkok Renewable Energy's head office is located in Bangkok, at SV City, 23th **floor, Rama** III Road.

Figure 8. An overview of the Bangkok Renewable Energy home page



- **Sustainable Power Corp**
- **News & Resources**
- **The Rivera Process**
- Sustainable Power Corp. is an international green energy total service provider focused on **environmentally safe biofuels manufacturing** and power generation.
- **SSTP** is now looking at expansion by placing factories into Central America, Sweden, and the Pacific Northwest of the United States.
- Even one fully operational SSTP factory in a **depressed country** can make a significant difference to the economy and standard of living in that country.

Figure 9. AView from iWISE of the Sustainable Power Corp home page

Based on Figures 4 to 9, a user is able to gain an overview and requires to spend less time on each Web page as they can be used as a summary. Using the DVT and information from the tag cloud and summary, a user may determine whether the site is appropriate to meet their needs. Moreover, the Document Type View on each page represents the complexity of a page as each node uses a unique color to represent a HTML tag. For example, if there are many nodes that represent paragraph tag, it means that there is more text content. Similarly, the text summarization and tag cloud will provide a quick summary of the key words and phrases contained in the sites. However, it is recognized at this stage the security system incorporated in the site is limiting the capability of this approach at this moment.

5. CONCLUSION

The World Wide Web (WWW) has become a huge repository of information and it also serves as an essential channel for information exchange. There are already in existence many hundreds and thousands of web pages of information on distributed power generation and alternate energy

[4] http://www.scottishpowerrenewables.com
[5] https://www.clpgroup.com
[6] http://www.bkkre.com
[7] http://www.sustainablepowercorp.us

development. Although web pages with such information can be accessed and ranked by search engines, there is no indication on the amount and the diversity of content in any of these pages. The users have to access the Web page and navigate through the site before they can determine whether the information is appropriate. This paper proposed an interactive summarization framework called iWISE to facilitate the process. The proposed approach is based on Document type view (DTV), Text Summarization and Tag Cloud. In addition, thumbnails associated with specific nodes can be provided in drill-down mode. Four web sites from companies or organizations dealing with power generation and alternate energy development have been compared. The next phase of the research is to provide a quantitative and qualitative assessment based on the summarization. It is believed that the approach will assist researchers to acquire relevant information and be able to carry out objective assessment of the websites.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

[1] EIA, *Annual Energy Review 2007*. 2008, Energy Information Administration.

[2] EIA. *World Energy Overview: 1995-2006*. 2008 June-December 2008 [cited 2009 Mar]; Available from: http://www.eia.doe.gov/iea/overview.html.

[3] Davis, Stacy C., Susan W. Diegel and Robert G. Boundy, *Transportation Energy Data Book: Edition 27*. 27 ed. 2008, Oak Ridge. 361.

[4] EIA, *Annual Energy Outlook 2009 with Projections to 2030*. 2009, Energy Information Administration.

[5] Berrah, Noureddine, et al., *Sustainable energy in China: the closing window of opportunity*. 2007, Washingtom: The World Bank.

[6] Cubasch, Ulrich, Yihui Ding, Cecilie Mauritzen, Abdalah Mokssit, Thomas Peterson , Michael Prather, *Historical Overview of Climate Change Science*. 2008, The University Corporation for Atmospheric Research: Colorada. p. 36.

[7] Evans, Robert L., *Fueling Our Future : An Introduction to Sustainable Energy*. 1st ed. 2007, Cambridge: Cambridge university press.

[8] RoughGuides, *The mini rough guide to Energy and our planet*. The first edition ed. 2008, London: Rough Guides. 96.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[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.

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

[20] 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.