International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)

International Conference on Information Systems and Computing (ICISC-2013), INDIA.

SURVEY ON WEB PAGE VISUAL SUMMARIZATION

A. Porselvi[1], S. Gunasundari[2]

[1]*Student, Department Of CSE, Velammal Engineering College, Chennai, India*
[2]*Assistant Professor – II , Department Of CSE, Velammal Engineering College, Chennai, India*
Email: porselviia@gmail.com / gunaannauniv@gmail.com

Abstract

Information on World Wide Web is congested with large amount of documents. Web page summarizations have received much attention in web intelligence, aimed to find relevant result and make better understand of result web page. The result web page contains a document title, a snippet, a URL. The snippet is a short summary of the document, which is designed so as to allow the user to decide its relevance. Text summarization can be classified into extractive and abstractive summarization. A visual summarization web page is efficient approach to identify relevant contents.

The picture in web page could be providing contextual description because the picture is worth a thousand words. Which including internal image, visual snippet, Thumbnails, External images. Thumbnails are scaled down image, which is mainly proposed for re-finding task. Visual snippet contains title information, the dominant internal image, and logo. Internal images can be directly used to summarize the web page. Although many web pages do not contain internal image, this problem is solved by External image. External images are searched and retrieved from the entire Internet. The more number of web page abstraction has been based on textual representation then visual representation. The extensive survey paper aims to analyze various type of web page summarization and also address the importance of visual web page summarization.

Keywords **– Text summarization, Visual Summarization, Internal Image, External Image, Visual Snippet, Textual Snippet.**

I. INTRODUCTION

Web Page Summarization is programs that search documents for specific query and returns a compact summary for a given web page to representing its main content. The goal of summarization is to produce coherent summaries that are good as human-authored summaries. Textual snippet is the most widespread search-based summarization. When a user submitted a query, web search engine provide the reference for sequence of top-k documents. Each document contains a title, a snippet, a URL.

Visual Summarization is an attractive new scheme to summarize web pages, which can help achieve a friendlier user experience in search and re-finding tasks by allowing users quickly get the idea on what the web page is about and helping users to recall the visited web pages. It provides a reliable summarization on web pages with dominant images Textual summarization has the disadvantage of users spend more time to read long snippet.

Short snippet doesn't convey enough information. Though visual summarization address this problem to some extent, web pages are visually summarized, since it is easier for people to get a quick understanding by seeing an image even without reading the textual snippet. This paper is organized as follows. In section II deals with web page summarization. In Section III, we deal with textual summarization.

Section IV briefly reviews visual summarization. At last, we conclude this paper in section V.

II. WEB PAGE SUMMARIZATION

Web page summarization help users get an idea of the page contents without having to spend time browsing the sites. This extracts the most important sentences of a web page and provides a summary to the user. The web includes different kind of information like text, images, video and audio. So we need to extract relevant result. The good web page summary must be a clear, a simple guide what is on the page.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

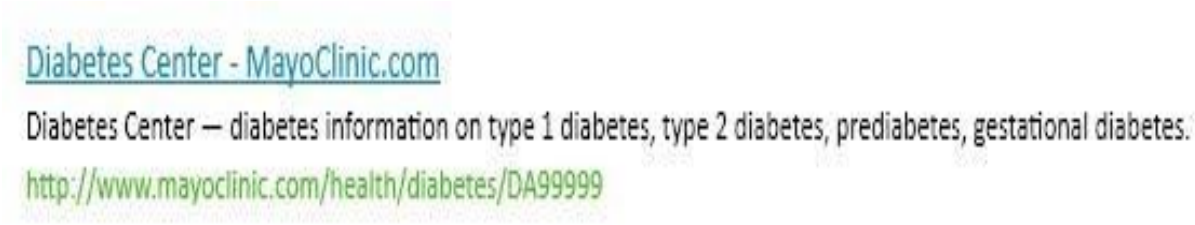

Fig1: Textual Snippet (This picture is taken from Jaime et al [8])

Which is must be useful in search result. The automatically summarized web page is achieved by two technique which are machine learning and natural language processing technique. Textual summarization extracts the relevant text or sentence from the web page. Image based summarization (or abstraction) of a website is the process of extracting the most important images from it.

Adam et al [16] suggest web page summarization using dynamic content, generally two type of web page summarization are content and context based methods. Both methods consider only fixed contents and characteristics, not account of their dynamic nature. This approach is towards automatic

III. TEXTUAL SUMMARIZATION

Search engine typically represent web pages in that result lists as textual snippet, with a title, a query-biased page summary (snippet), and a URL. Fig 1 shows example of text snippet. Text summarization is a tool to summarize large document of text. It is very difficult to manually summarize large document. It is a brief but accurate representation of the contents of document.

Thakkar et al [4] suggest text summarization by graph based algorithm. Which presents innovative unsupervised methods for automatic sentence extraction using graph-based algorithms and shortest path algorithm .which has advantage of visualizing a large text document within a short duration. The purpose of providing a text summary for each result page is to enable the user to quickly judge whether it is what he or she needs.

Karel et al [13] suggest automatic text summarization based on algebraic reduction methods. Their strong properties are that they do not depend on A particular language.

The goal of automatic summarization is compress the source text into a shorter version preserving its main content.The most widely used algebraic reduction methods are latent semantic analysis (LSA), Non - negative matrix factorization (NMF). LSA is a fully automotive algebraic-statistical technique for extracting and representing the contextual usage of words Jaime et al [8] suggest text snippet contains a title, a one line summary, a URL. The display of text is not specific for particular user query and also which doesn't exist of hint highlighting.

Summary evaluation [14, 15] is a very important aspect for text summarization. Summaries can be ranked using intrinsic or extrinsic measures. Intrinsic measures to automatically determine the quality of a summary by comparing it to other summaries created by humans. They introduce four different ROUGH measures, ROUGE-N, ROUGE-N, ROUGE-W and ROUGE-S. ROUGE an automatic evaluation package for summarization. The pyramid method assigns a score to summary and allows the investigator to find what important information is missing, and thus can be directly used to improvements of summarization. However textual summarization is differ from textual snippet, Textual snippet is a one line short summary of document, whereas summarization is the process of generating web abstraction result page.

IV. VISUAL SUMMARIZATION

4.1 Internal Image

The concept of image-based summarization for improving the quality of website summaries and as a tool for more effective web browsing and retrieved. Internal Image [3, 5, 7, 8, 9, 10] based visual summarizations are commonly adopted in existing products. Li et al [7] proposes internal

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

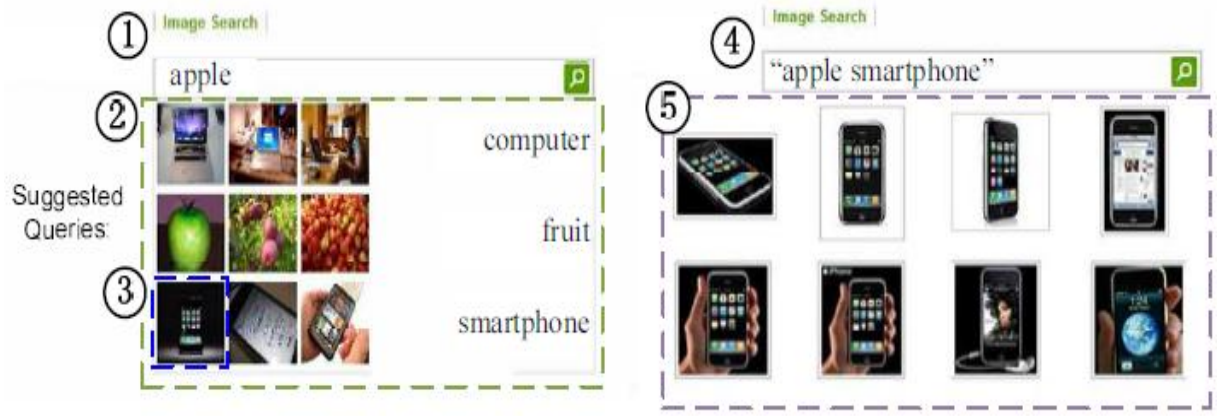

Fig 2: The workflow of a VQS system. A user can interactively: (1) Submit a keyword query (2 Browse keyword – image suggestion provided by VQS system (3) Select on suggestion to specify the search intent (4) Expands original query with corresponding keyword (5) Performs image search based on new query (4) (This picture is taken from Zheng et al [6]).

Images based technique which extracts the dominant Images from the web pages as "image excerpt". The dominant images are first deducted by a trained model based on three levels of image features. And then the most relevant images are selected based on the query and the image's surrounding text.

Many popular search engines [Google, Yahoo!] have developed technologies that allow users to search web images. However as an aforementioned, for a large amount of web pages, dominant images are unavailable, which limits the applicability of search approaches.

Internal image summarization [6, 11,] is reasonable because dominant images can present the ideas of the web pages and often they are impressive to attract user's attention. Zhen et al [6] suggest automatically provide a list of textual query terms based on user's current input query, which can be called as Textual-query suggestion. Which provides a more efficient query interface to formulate an intent-specific query by simultaneously provides both keyword and image suggestions. This can be able to help user's specify and deliver their search intents in a more precise and efficient way, which perform image search using text search technique. Fig 2 shows the process of query suggestion for image search, which has five steps to produce result. Baratis et al [9] suggest image based summarization, which focus on logo and trademark images. Their system incorporates machine learning for distinguish logo and trademark from images.

Duplicate logo and trademark images are deleted and only unique logo and trademark images are extracting. The experimental result of 50 web sites shows that image summarization reached up to 64% accuracy. Feature work includes experimental with larger training data sets and image types for improving the performance machine learning.

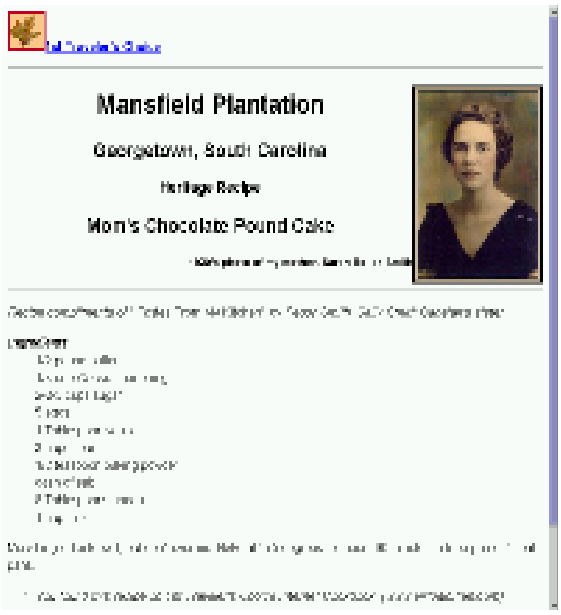

Fig 3: Plain Thumbnails (This picture is taken from Woodruff et al [11])

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

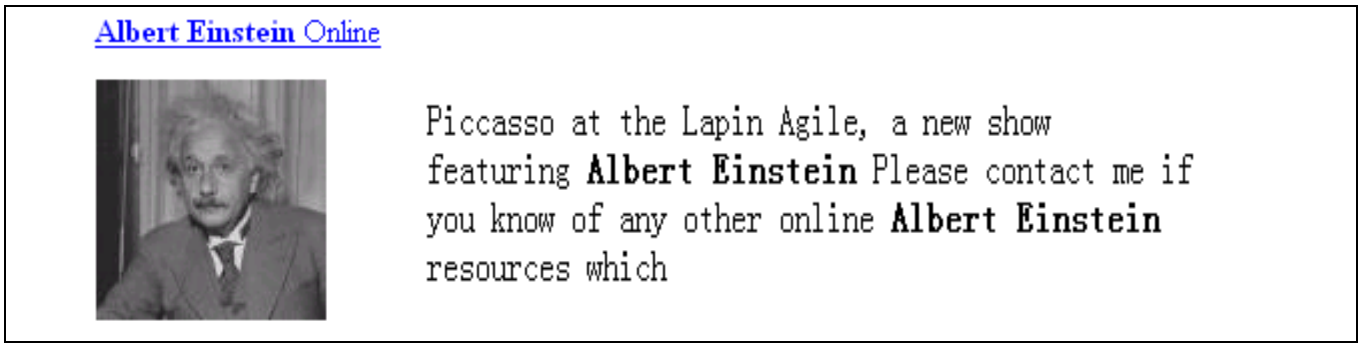

Fig 4: Search result of query Albert Einstein (This picture is taken from Zhiwei et al [7])

Binxing et al [3] suggest internal image is one of the most widely used web page summarization method. The summary of web page is described by the image contained in those web pages, which is directly used to summarize the web page. The web page has lot of images, including logo, advertisement etc. We need to deduct most dominant image among all of them, for this dominant image deduction we used learning based algorithm. Image level, page level, site level are the three feature of dominant image detection. Since there are multiple dominance levels, so they consider dominance detection issue is a ranking problem. They used linear Ranking - SVM to train the ranking model for dominant detection.

4.2 Thumbnails

The most common web page summarization is a scaled down bitmap or thumbnails [7, 11, 12, 3].which displaying a snapshot of a particular web page as rendered in the browser. While thumbnails approach are mainly uses for re-finding task. The number of studies [5, 11,] have involved thumbnails based summarization .although thumbnails are perceived as images, people usually need to read textual information presented in thumbnail previews. This has disadvantage of additional time cost to download and reading difficulties due to poor accessibility of textual information presented in thumbnail previews.

This has disadvantage of additional time cost to download and reading difficulties due to poor accessibility of textual information on thumbnails. Thus woodruff et al [12] proposed textually enhanced thumbnails, which enhance the readability of certain parts of the document within the thumbnails and display highlighted keywords transparently overlaid on the reduced document. Fig 5 shows the textually enhanced thumbnails, the extracted lines are highlighted by color letter. The enhanced thumbnail system works in the three stages, HTML modification, rendering, image modification. This can address the raw thumbnails problem by Some extent. Zhiwei et al [7] presents the thumbnails of web page for web search, which suggest that thumbnails of web pages used along with text snippets in search engine interface. It helps users reducing Predicting errors, at a little time cost in processing time, which help uses, make quicker relevance judgment of search result for a wide range of search queries. Fig 4 shows the thumbnails with textual summary. By viewing thumbnails and reading text users can have better understand of result.

4.3 Visual Snippet

Jaime et al [5] suggest visual snippet is a recently proposed web page summarization approach which enriches internal dominant images by Fig 6 shows the example of visual Snippet of web page.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

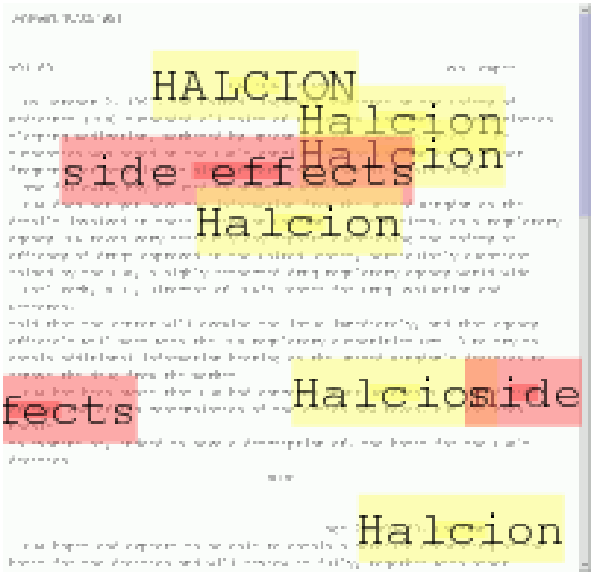

Fig 5: Textually enhanced thumbnails (This picture is taken from woodruff et al [12])

Fig 6: Web Page (Right) and the visual snippet build from that web page (Left) (This picture is taken from Jaime et al [8])

Text snippet is effective for finding new web pages that have never seen before. Thumbnails, in contrast good for supporting re-finding task, a visual snippet captures the better of these two representations. They use the two components identified through design analysis to build small representations for an arbitrary web page. (a) identified the component pieces (b) compiling the component pieces. Visual snippet can explore how text summaries, raw thumbnails and visual snippet perform in both search and re-finding tasks.

4.4 External Images

Internal image based web page summarization are adopted by industrial product, such as Google's "Site with images". However for a large amount of pages do not contain internal images. Beside these problems are solved by "External Images". External Images is a new approach for web page summarization. This is used to retrieve relevant images from the Internet. So every summarization result page contains relevant image, User can get quick and better understand by just seeing image.

Which saves their valuable time without spend time to read irrelevant result. External image based summarization is newly adopted technique, which are not adopted by existing product? External images are reasonable because which are impressive to attract user attention.

Binxing et al [3] presents visual summarization of web page, to generate meaningful Visual summarization for those web pages doesn't exist of internal image. The innovative idea is external image. The following steps used to produce external images are, key phrase extracted, image search based on key phrase, Re-ranked the returned image, finally the top ranked images are used to summarize the target web page. The external images are out performed the internal images. However which is suffering from reliability problem. The early system used a straightforward strategy is to select external Images when the internal image threshold value is Below a predefined threshold, however the threshold value is hard to determine therefore the clustering based algorithm is used for jointly select the best summarization from both of internal and external images.

V. CONCLUSIONS

In this paper, surveys of various kinds of web page summarization methods have been presented Based on the above summary; the visually summarizing web page work is less then textual summarization. Image based summarization could lead to more comprehensive summaries and allow for more effective Web browsing and retrieval. Whereas, in existing systems of visual methods used thumbnails and internal images. Web pages has complex layout and rich content, users are difficult to see clearly anything from Small thumbnails, although many of web pages do not contains on internal images. Those problems are extremely solved by external images. External image are not reliable, which have the property of availability, so we use internal images which is more reliable but doesn't has availability property. These types of Internal and external images jointly taken into consideration for summarizing web pages, since they have respective advantages and may complement each other. Many research need to be done in Visual summarization of web pages.

REFERENCES

[1] Omara, F.A, Amoon, M, E1-Fishawy, N.A; E1-Kazaz.S, "Analyzing anchor links to enhance the web-snippet clustering technique" Informatics and Systems(INFOS),8th conference,Cairo,2012,pp7-11.

[2] Ragatha, D.V, Yadav, D. "Image Query Based Search Engine Using Image Content Retrieval" Computer Modeling and Simulation (UKSim), Cambridge, 2012, pp: 283-286.

[3] B. Jiao, L. Yang, J. Xu and F. Wu, "Visual summarization of web pages," in SIGIR '10.New York, NY, USA: ACM, 2010, pp.499-506.

[4] Thakkar, K.S, Dharaskar, R.V, Chandak, M.B, "Graph based algorithm for Text Summarization "Emerging in Engineering and Technology (ICETET), 2010 3rd conference, Goa, India, pp.516-519.

[5] Jaime Teevan, Edward Cutrell, Danyel Fisher, Steven M.Drucker, Gonzalo Ramos, Paul Andre, Chang Hu, "Visual Snippets: Summarizing Web Pages for Search and Re-visitation" CH '09: Proceedings of the 27th international conference on Human factors in computing systems, pages 2023-2032, New York, NY, USA, 2009.ACM.

[6] Z-J. Zha, L. Yang, T. Mei, M.Wang and Z. Wang, "Visual query suggestion," in proceedings of the seventeen ACM International Conferences on Multimedia. New York, NY, USA: ACM, 2009, pp. 15-24.

[7] Z. Li and L. Zhang. "Improving relevance judgment of web search results with image excerpts" In WWW '08: Proceedings of the 17th international conference on World Wide Web, pages 21-30, April 2008.

[8] Mei Wang, Honatao Xu, Guoyu Hao, Xiangdong Zhou, Wei Wang, Qi Zhang, Barlie Shi, "Picture Book: A Text and Image Summary System for Web Search result" Data Engineering, ICDE 24th International conference,2008, Pages 1612-1615.

[9] Baratis, E., Petrakis, E.G.M., Millios, E., "Automatic Website Summarization by Image Content: A Case Study with Logo and Trademark images," Knowledge and Data Engineering, IEEE. Pages 1195 – 1204, September 2008.

[10] Zhouyao Chen, Ou Wu, Mingliang zhu, Weiming Hu, "A Novel Web Page Filtering System by Combining Text and Images" Web Intelligence, WI 2006, IEEE/WIC/ACM International Conference, pages 732-735, Hong Kong, China, 2006.

[11] Woodruff, A., Faulring, A., Rosenholtz, R., Morrison, J., and Pirolli, P. (2001). "Using Thumbnails to Search the Web", In Proceedings of CHI '01, Pages: 198-205.

[12] Woodruff, A., Rosenholtz, R., Morrison, J., Faulring, A. and Pirolli, P. (2002). "A comparison on the use of text summaries, plain thumbnails, and enhanced thumbnails for Web search tasks. JASIST, 53(2), 172-185.

[13] Karel Jezek and Josef Steinberger, "Automatic Text Summarization", Vacalv snasel(Ed.): Znalosti 2008,pp 1-12, FIIT STU Brarislava, Ustav Informatiky a Softveroveho inzinierstva,2008.

[14] Ani Nenkova and, Rebecca Passonneau, "Evaluating Content Selection in Summarization: The Pyramid Metho" In HLT-NACCL, pp 145-152, 2004.

[15] Chin-Yew Lin " ROUGE- A Package for Automatic Evaluation of Summaries" in Proc. ACL Workshop on text summarization branches out, 2004.

[16] Adam Jatowt, "Web Page Summarization using Dynamic Content," in proceedings of 13th international World Wide Web Conference on alternate track papers and posters, ACM, New York, NY, USA, pages 344 – 345, 2004.