

A Quantum Based Web Summarizer for Children's News Rendering

Enikuomelin A. O. *, Rahman M. A.

Dept, of Computer Science, Lagos State University

Abstract Dissemination of information to children has often been an issue of concern to web services consultants. Such is predetermined by the complexity that exists between managing the web content for children's view and sharing such useful content among them. Particularly, the process of dissemination of opinion, browsing through archived reviews to locate different opinions on particular topics, is a time-consuming and tedious task, and in most cases, the large amount of available information makes it difficult for users to absorb. To facilitate the process of synthesizing opinions expressed in various web domains essentially, the news portal, on a particular topic specified in a user query/question or as contained on the web, a quantum-based multi-document opinion summarizer is introduced. This creates a summary in response to queries or already available information, which either reflects general opinions on product, concept or opinion as specified in the query, by (i) identifying facets discussed in the reviews retrieved in response to query and (ii) employing a sentence-based, opinion classifier to determine the polarity of each sentence in each review. These steps dictate which sentences are included in the summary. The paper explores the processes of quantum based news rendering and show that such approach is useful in the presentation of news opinion for children understanding. The paper concludes by showing the usefulness of summarization to the general of Information Retrieval process.

Keywords Web, News Rendering, Information Retrieval, Summarization

1. Introduction

With rapid development of Internet Technologies and Web Explosion, searching useful information from huge amount of Web pages becomes an extremely difficult task. Currently, Internet search engines are the most important tools for Web information acquisition. Based on techniques such as Web page content analysis, linkage analysis, etc., search engines locate a collection of related Web pages with relevance rankings according to user's query. However, current search results usually contain large amount of Web pages, or are with unintuitive rankings, which makes it inconvenient for users to find the information they need. Therefore, techniques for improving the organization and presentation of the search results have recently attracted a lot of research interest. The typical techniques for reorganizing search results include Web page clustering, document summarization, relevant information extraction, search result visualization, etc.

Children news portal is a special kind of news website that is specially made for children for the use of information on the internet; it basically does not allow its usage to children

of all ages but to those that are competent with the use of computers to explore the internet. Some information, that such portal provides to children includes:

- Date, time and venues of recreational activities
- Homework help and references
- Online uploaded files such as games, music, books, etc.

All these information, most times, come in large, voluminous text files and as documents on web pages. In order to save time and relief children from the stress of reading those large voluminous text files and document, summaries of these information are provided by a web-based tool known as "web summarizer".

The web summarizer can summarize word documents, web pages, PDF files, email messages and even text from the clipboard.

The concept of a web summarizer for the development of children news portal emerged as a result of the difficulties experienced by children in the retrieval of relevant information from large, voluminous text files and documents on web pages. By focusing on the relevant key sentences contained within a document, web summarizer enables children to browse quickly through volume of information and extracts the documents most applicable to their search requirements.

The volume of reviews for particular news can often be prohibitive and time consuming for potential users/children that wishes to read all relevant information, compare

* Corresponding author:

toyinenikuomelin@gmail.com (Enikuomelin A. O.)

Published online at <http://journal.sapub.org/ijit>

Copyright © 2015 Scientific & Academic Publishing. All Rights Reserved

alternatives, and make informed decision. Thus, the ability to analyze a set of online reviews and produce an easy to digest summary is a major challenge for children in the retrieval of relevant information on their news portal. Web summarizer therefore helps develop children news portal by utilizing consistent sentence selection criteria that matches the conceptual content of documents.

A web-based summarization system that helps develop children news portal, as explained, implores the use of the internet to take as input, a set of reviews for specific news and produces a set of relevant (feedback) aspects, an aggregate specification for each aspect and supporting textual evidence. However, there are some problems in the course of summarization. Examples of these problems are as follows: -

- Client/servers problem: being a web system, request sent by clients (i.e children) for summarization and the summarized texts/documents from web server (web summarizer) often take time to deliver due to web traffic [8]. This is known as response time problem.
- Presentation problems: problems sometimes occur in the presentation of summarized text/document. Such problems include the following:
 - i. Linguistic problem: this problem arises when summaries are not presented in language children desires.
 - ii. Threshold/number of lines problem: the desired number of lines for presentation of summarized text/document sometimes may not present 100% relevant information. More number of lines may be required.
- Complex problem: problem experienced as a result of eliminations of some information and a simultaneous introduction of new ones makes summarization process so complex for children on their news portal.

Text Summarization is an active field of research in both the Information Retrieval, IR, and Natural Language Processing, NLP communities. Summarization is important for IR since it is a means to provide access to large repositories of data in an efficient way. It shares some basic techniques with indexing, since both indexing and summarization are concerned with identifying the essence of a document.

In comparison with IR, the field of summarization suffers from the difficulty of defining a well specified and manageable task. Since truly reusable resources like the TREC test collections did not exist for summarization, it was hard to measure progress. Importantly, the high amount of variance across human summaries complicates evaluations.

2. Background

The importance of understanding the function a summary serves for users is widely acknowledged, and seminal works defining summary types by functions [1], [2] are frequently

cited by researchers. Task orientation defines extrinsic technology assessments, and the research literature on how to assess performance for machine generated summaries in an experimental task scenario has grown [3], [4].

An increasing number of research papers on summarization systems also describe some type of extrinsic evaluative task [5], [6]. A number of factors (i.e. characteristics of summaries, documents, users, and tasks) have surfaced which have implications for technology use. More research assessing technology (or any aspect of it) in-use on a user's own data even in a development mode along the lines of [8] is needed. While experimentation designs involving subjects performing short term controlled tasks may yield results of statistical significance, generalization to the user community is limited. In addition, the level of user support text summarization systems should provide and also continues to be speculative. More interest lies in new areas of inquiry like visualization and browsing techniques [9], multi-document summarization [10-12], multi-media summarization [13].

3. Technology-Related Modifications

A. User-Centered Changes to Technology Work Practices
On technology performance, it was understood that

- seamless integration with an IR system was preferred
- users with static queries were more likely customers for a summary service
- gains in efficiency are hard to measure for a task already efficiently performed in a real-world situations.

In response, a summary service was established in which retrieval results are directly routed to the summary server and await the user. Integrating the summarization tool into the IR system was planned (Uploading batches and then submission to the server is still an option.) Another thing that was abandoned was the naive idea that data overload equates to summarization requirements and realized that the technology does not apply to all users. In order to demonstrate tool summarization efficiency, the base lining of full-text review is needed. A number of options were considered, but rejected; user self-report and timing, observations, and even the creation of a viewing tool to monitor and document full text review. Instead, developers baselined full text scanning through information retrieval logs for a subgroup of users by tracking per document viewing time for a month period. These users submit the same queries daily and view their documents through the IR system browser. For the heaviest system users, 75% of the documents were viewed in under 20 seconds per document, but note that users vary widely with a tendency to spend a much longer browse time on a relatively small number of documents. Then subgroups of these users were identified and developers attempted to deploy the summarizer to this baseline group to compare scanning time required over a similar time frame. These data are currently analyzed.

System in a work environment is considered a good indicator of tool utility, but developers wanted some gauge of summary quality and also anticipated user concerns about an emerging technology like automatic text summarization. Developers compromised and selected a method to measure the effectiveness of their summaries that serves a dual purpose: their users gain confidence in the utility of the summaries and they can collect and measure the effectiveness of the generic summaries for some of the users on their data. Developers initially piloted and now have incorporated a data collection procedure into their application. In the on-line training, they guide users to explore tool capabilities through a series of experiments or tasks. In the first of these tasks, a user is asked to submit a batch for summarization, then for each of five to seven user-selected summaries to record answers to the question:

"Is this document likely to be relevant to me?" (based on the summary)

__yes __no

Then, the user was directed to open the original documents for each of the summaries and record answers to the question:

"Is the document relevant to me?" (after reading the original text)

__yes __no

In a prototype collection effort, users were asked to review the first ten documents, but in follow-on interviews the users recommended review of fewer documents. They understand the limits this places on interpreting their data. Also, the on-line training is optional so they are not able to collect these data for all their users uniformly.

4. Analysis of the Proposed QMOS (Query-Based Multi-Document Opinion Summarizer)

This section detailed the processing steps of QMOS in creating a summary of multiple reviews retrieved in response to a user's query. The overall process of QMOS is illustrated in Figure 1.

The design methodology of each individual component of QMOS addresses a particular research problem on its own, which include:

- (i) identifying products, facets, and opinionkeywords in a user's question using a multi-class SVM on a number of novel features,
- (ii) finding opinions on various facets of product P using a novel sentence clustering algorithm based on word correlation factors,
- (iii) condensing each individual review to exclude sentences in the review that are redundant, relatively uninformative, or lack of opinions, and
- (iv) ensuring that each QMOS-generated multidocument summary is non-redundant, coherent, and concise by employing a simple, yet highly effective, sentence

selection algorithm.

4.1. Identifying the User's Information Needs

QMOS adopts a one-against-all implementation of a multiclass SVM to identify information needs expressed in a query.

To train a multi-class SVM, each training instance is an input vector of a non-numerical, non-stopword K in a query Q and is a succession of '1' ('0', respectively), each of which represents the presence (absence, respectively) of an SVM-feature F if F applies (does not apply, respectively) to K.

- **Is-Singular** is set (to '1') if K is in a singular form. *Products*, *opinionkeywords*, and *non-essential terms* tend to be expressed in singular form.
- **Is-Capitalized** is set if the first letter of K is capitalized. The first character of a *product* is often capitalized.
- **Is-Adjective** is set if K is given an *adjective* part-of-speech (POS) tag. QMOS employs the Stanford POS tagger for assigning POS tags, such as noun, verb, or adjective, to keywords (in Q). *opinionkeywords* specified in Q are often adjectives which describe different aspects of a product in Q.
- **Is- opinionis** set if K is a *opinionkeyword*, which is determined by using a list of more than 4,000 opinionkeywords provided by the General Inquirer.
- **Is-After-Preposition** is set if K appears immediately after a preposition, which is identified using *Stanford-POS*. Both *products* and *facets* tend to occur after a preposition in Q.
- **Is-After-Apostrophe** is set if K appears immediately after a term in a Saxon genitive form, i.e., a traditional term for the apostrophe-s. *Facets* often appear after a term in the Saxon genitive form in Q.
- **Is-Before-opinionis** set if K appears immediately before a opinionkeyword in Q. Both *products* and *facets* are often followed by a opinionkeyword in Q.
- **Is-Stopword** is set if K is a stopwords, which is a *nonessential* term.
- **Is-5W1H** is set if K is one of the keywords frequently used in formulating questions, i.e., "what", "when", "where", "who", "why", and "how". 5W1H terms are treated as *nonessential* terms, since "when", "where", "who", and "why" do not appear often in opinionquestions, whereas "how" and "what", which appear more often, do not have a direct impact on the information needs specified in users' questions.

QMOS also detects *negation* terms (or their stemmed versions) in Q, such as "not", "no", "without", "except", "excluding", "remove", "nothing", and "leave out", so that the polarity of a *opinionkeyword* in Q is *reversed* if it is *preceded* by a *negation* term.

EXAMPLE 1 shows a sample user query, denoted QP, along with the information needs in QP identified by using keyword types in the multi-class SVM of QMOS.

Query Qp: What do people like about Twilight's story?

Products: "Twilight"
 Facets: "Story"
 opinionKeywords: "Like"
 Non-essential Terms: "What", "Do",
 "People", "About"

4.2. Retrieving Relevant Reviews

Having identified the product P specified in a query Q, QMOS queries retrieve reviews on P for creating the summary for Q. Since search engines at these websites simply process exact keyword searches without accurately relating the opinion applied to P in Q, QMOS extracts reviews from these websites using a simplified query Q' which includes solely P specified in Q. QMOS gathers the top reviews extracted by review repositories in response to Q', which yields the set of reviews R for creating a summary (of R) in response to Q.

F. Generating Summaries

QMOS creates a single summary in response to the information needs specified in a user query Q. A summary is:

- (i) *General*, if Q inquires on common feedback of a particular product P,
- (ii) *Sentiment-Specific*, if Q asks for positive or negative information about P,
- (iii) *Facet-Specific*, if Q queries on specific facets of P, or
- (iv) *Facet-Sentiment-Specific*, if Q looks for opinion information on specific facets of P.

A length of (approximately) 250 words is adopted as the size of each QMOS-generated summary according to the guidelines defined for (some of) the datasets archived by the Text Analysis Conference (TAC), which provides benchmark datasets for assessing the performance of a query-based multi-document summarizer. Since QMOS is a sentence-based, extractive summarizer, QMOS includes in a summary as many sentences as necessary up till the size limit so that the total number of words in the summary being constructed after including the second to the last chosen sentence is less than 250..

G. Evaluating the Quality of QMOS-Generated Summaries

Following the evaluation methodology established by TAC, this work relied on independent appraisers, which manually evaluated the *grammaticality*, *non-redundancy*, *structure and coherency*, *responsiveness*, and *readability* of the summaries created by QMOS using TAC-08.

To establish a baseline measure on the quality scores achieved by QMOS-created summaries, I compare the performance, based on quality scores, of some automatic multi-document summarizers with QMOS based on summaries created

In fact, there is *not* a single system examined that achieves higher scores than QMOS in all of the quality measures. QMOS achieves the highest *responsiveness* score among all the summarizers and can only be outperformed by any summarizers in creating summaries that are *non-redundant*. The summarization systems that outperform QMOS on either the *grammar*, *readability*, or *structure and coherency* quality measures employ natural language processing techniques to touch up the summaries shown to users as the final products. The summarization approach of QMOS, on the other hand, is purely extractive i.e., it solely extracts sentences in the original document(s) without refinement to create a summary. The summarizer which achieves a better *non-redundancy* quality measure than QMOS uses textual graphs to model opinions in a review and relies on word order and part-of-speech tags to determine redundant sentences, neither of which is employed by QMOS for simplicity.

5. Detailed Implementation Description

The implementation of the system designed in this project work serve as a perfect means of summarizing results of web pages. The system is a search engine (that helps in children news portal) which retrieve images, results of search and the summarized texts of submitted queries. Images (with limited amount) are stored in the database of our system and can be accessed at any point in time by search using a keyword. A sample of such output as a search result is shown in figure 1 below.

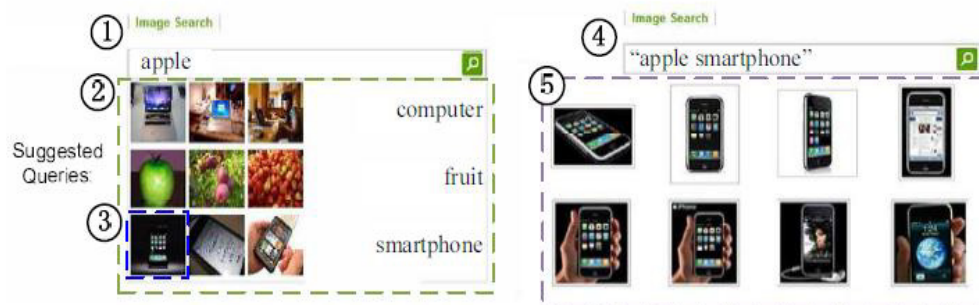


Figure 1. Sample Image Output

6. Summary and Conclusions

During the last decade, IR has been changed from a field for information specialists to a field for everyone. The transition to electronic publishing and dissemination (the Web) when combined with large-scale IR has truly put the world at our fingertips through the process of computerized summarization. This paper considers the challenges of this wide acceptability essentially as regards information usage by children and thus proposed the QMOS as a summarization technique for news rendering for children. Elaborate technique discussion was carried out and theoretical experiments shows that the use of QMOS is indeed a shift from the past. This area will further be explored as a means of easing the use of web information.

REFERENCES

- [1] Aone, C., Gortalsky, J. and Okunowski, M.E. 1997. Trainable, scalable summarization using robust NLP. In *Intelligent Scalable Text Summarization*. Madrid, Spain: Association of Computational Linguistics.
- [2] Boguraev, B., Kennedy, C., Bellamey, R., Brawer, S., Wong, Y.Y. and Swartz, J. 1998. Dynamic presentation of document content for rapid on-line skimming. *Intelligent Text Summarization*. (Papers from the 1998 AAAI Spring Symposium Technical Report SS-98-06). Cambridge, Massachusetts: MIT Press.
- [3] Futrelle, R. 1998. Summarization of documents that include graphics. *Intelligent Text Summarization*. (Papers from the 1998 AAAI Spring Symposium Technical Report SS-98-06).
- [4] Morris, A., Kasper, G., and Adams, D. 1999. The effects and limitations of automated text condensing on reading comprehension performance. In Mani, I. and Maybury, M. (eds.), *Advances in Automatic Text Summarization*. pages 305-323.
- [5] Salton, G., Singhal, A., Mitra, M. and Buckley, C. 1999. Automatic text structuring and summarization. In Mani, I. and Maybury, M. (eds.), *Advances in Automatic Text Summarization*. pages 342- : 355, Cambridge, Massachusetts: MIT Press.
- [6] Strzalkowski, T., Wang, J. and Wise, B., 1998. A robust practical text summarization. *Intelligent Text Summarization*. (Papers from the 1998 AAAI Spring Symposium Technical Report SS-98-06).
- [7] Dragomir R. Radev, Kathleen R. McKeown, (1998), Generating natural language summaries from multiple on-line sources. "Computational linguistics", 24(3): 469-500, September.
- [8] Enikuomehin T, (2013), "Web Development Lesson note", In LNCS, Unilorin.
- [9] Hornby A. S, (2005), "Oxford ADVANCED LEARNER'S Dictionary" 7th Edition. Sasha Blair- Goldensohn, (2008), "Google submit".
- [10] Attardi, G. A. Gulli, and F. Sebastiani. Automatic Web PageCategorization by Link and Context Analysis. In Chris Hutchison and Gaetano Lanzarone (eds.), Proc. of THAI'99,1999, 105-119.
- [11] McKeown, K. Jordan, D. and Hatzivassiloglou, V. 1998. Generating patient specific summaries on online literature. *Intelligent Text Summarization*. (Papers from the 1998' AAAI Spring Symposium Technical Report SS-98-06).
- [12] Morris, A., Kasper, G., and Adams, D. 1999. The effects and limitations of automated text condensing on reading comprehension performance. In Mani, I. and Maybury, M. (eds.), *Advances in Automatic Text Summarization*. pages 305-323,
- [13] Salton, G., Singhal, A., Mitra, M. and Buckley, C. 1999. Automatic text structuring and summarization. In Mani, I. and Maybury, M. (eds.), *Advances in Automatic Text Summarization*. pages 342- : 355, Cambridge, Massachusetts: MIT Press.