Structured and Unstructured Document Summarization: Design of a Commercial Summarizer using Lexical Chains

Hassan Alam, Aman Kumar, Mikako Nakamura, Fuad Rahman[1], Yuliya Tarnikova and Che Wilcox
BCL Technologies Inc.
fuad@bcltechnologies.com

Abstract

The process of summarizing documents is becoming increasingly important in the light of recent advances in document creation/distribution technology, and the resulting influx of large numbers of documents in every day life. This paper presents a document summarizer that combines document analysis, structural decomposition, XML representation and lexical chain analysis. The proposed summarizer is compared to three commercially available summarizers and it is shown that it produces either comparable or better summaries overall.

1. Introduction

Document summarization has been a well-known field of computational linguists for many decades, but only recently has it been possible to commercialize this technology. The availability of affordable computers with very high memory and computing power is responsible for this. The computerization of our day-to-day life has resulted in easy access to documents and a reduction of privacy. These two factors are related, because as it is true that a paperless office has resulted in increased productivity and active cooperation and networking, it has opened a gate of unwanted information. Unsolicited email ("junk mail") is only a small incarnation of the problem. Unwanted information is now routinely passed on to people resulting in a deluge of documents, reducing productivity by wasting valuable time. This realization has created a demand for a technology that can filter or flag unwanted documents. While it is relatively simple to filter out unwanted emails from unknown sources by mapping keywords, sending addresses, topics etc., filtering out documents can be a completely different kettle of fish. A commercial summarizer will be very useful in this context.

This paper has proposed a new commercial summarizer using Natural Language Processing (NLP) techniques. The aim is to design a summarizer that not only processes traditional "flat" documents, which are primarily textual documents with no structure, but also to process complex structured documents by retaining the structure.

2. Background of summarization

Summarization is a widely researched problem. As a result, researchers have reported a rich collection of approaches for document summarization.

2.1 Academic approaches to summarization

There are two main types of resources available in the literature. The first is a class of approaches that deals with the problem of document classification from a theoretical point of view, making no assumption on the application of these approaches. These include statistical [1,2], analytical [3,4], information retrieval [5,6] and information fusion [7] approaches. The second class of resources deal with techniques that are focused on specific applications, such as baseball program summaries [8], clinical data visualization [9] and web browsing on handheld devices [10]. In addition, complete working systems have also been reported [11,12]. For a comprehensive review, the reader is referred to [13]. In general, these summarization techniques focus on the textual content of a document and the graphical or tabular information is largely ignored.

2.2 Commercial summarizers

There are some summarizers already commercially available in the market. They include Copernic® (http://www.copernic.com/index.html), Sinope® (http://-www.sinope.nl/en/sinope/index.html) and AutoSummarize, embedded as part of Microsoft® Word®. Copernic® produces summary reports for text contents by processing documents, web pages, hyperlinks, e-mail messages and files. Sinope®, generates summaries of arbitrary texts, including web pages, by integrating with Microsoft® Internet Explorer. AutoSummarize® allows

[1] Corresponding author. The authors gratefully acknowledge the support of a Small Business Innovation Research (SBIR) award under the supervision of US Army Communications-Electronics Command (CECOM).

summarization of Word® documents, but offers far fewer options. It only allows target specifications in terms of number of sentences, some percentages of size and some number of words, but does not allow any structural analysis, i.e. Table of Content (TOC)-type output.

3. The proposed document summarizer

The proposed document summarizer has multiple steps associated with it. The first is an analysis of the document structure. The second step is to classify the documents into a set of pre-defined categories. The third is to use natural language techniques to regenerate summaries of the textual content of the document. Finally, in the fourth step, the textual summaries are combined with the document structure extracted in the last step to generate the overall summary.

3.1 Document structure analysis

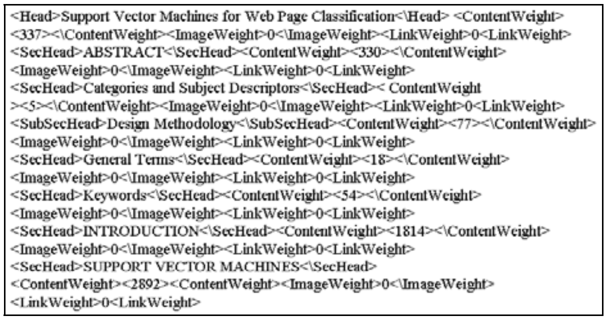
The structure of a document is defined in terms of headings, titles and sectional hierarchy. The principal attributes for detecting titles and section headings include font size, boldness, underline, and link properties. Once identified, heuristics are used to classify them as titles or section headings by analyzing their relative font size variations corresponding to other section headings and the surrounding text. This creates a hierarchy of sections and subsections ("Table of Content" or TOC, etc.), producing a structural summary of the document in terms of the sectional layout. This also provides information about the overall layout and content size of each section. Content may include text, images, links and other entries.

Figure 1 shows the extracted structural layout from the document of **Figure 3**. **Figure 2** shows the extracted structural layout from the document of **Figure 4.** The structural layout is described using a custom XML notation.

3.2 Document classification

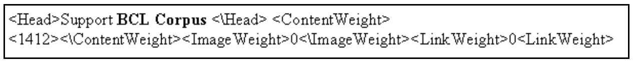
Documents are categorized into two classes, structured and non-structured. Structured documents have a well-defined hierarchical structure, such as titles and sections clearly marked with single or multiple level headings. Other attributes that create hierarchy, such as distinctive color, underlines, boldness, etc., are also considered. **Figure 3** shows an example of a structured document. A non-structured document (a "flat" document) will not have any of these attributes. These types of documents usually have a title, but after that the content is not organized in any structured fashion. **Figure 4** shows an example of a flat document.

Heuristics are used to classify a document in either of these two classes using the information gathered during the structural analysis. Once these attributes are detected and properly classified, it is easy to classify the documents into "structured" or "flat" categories.



```
<Head>Support Vector Machines for Web Page Classification<\Head> <ContentWeight>
<337><\ContentWeight><ImageWeight>0<\ImageWeight><LinkWeight>0<LinkWeight>
<SecHead>ABSTRACT<\SecHead><ContentWeight><330><\ContentWeight>
<ImageWeight>0<\ImageWeight><LinkWeight>0<LinkWeight>
<SecHead>Categories and Subject Descriptors<\SecHead>< ContentWeight
><5><\ContentWeight><ImageWeight>0<\ImageWeight><LinkWeight>0<LinkWeight>
<SubSecHead>Design Methodology<\SubSecHead><ContentWeight><77><\ContentWeight>
<ImageWeight>0<\ImageWeight><LinkWeight>0<LinkWeight>
<SecHead>General Terms<\SecHead><ContentWeight><18><\ContentWeight>
<ImageWeight>0<\ImageWeight><LinkWeight>0<LinkWeight>
<SecHead>Keywords<\SecHead><ContentWeight><54><\ContentWeight>
<ImageWeight>0<\ImageWeight><LinkWeight>0<LinkWeight>
<SecHead>INTRODUCTION<\SecHead><ContentWeight><1814><\ContentWeight>
<ImageWeight>0<\ImageWeight><LinkWeight>0<LinkWeight>
<SecHead>SUPPORT VECTOR MACHINES<\SecHead>
<ContentWeight><2892><ContentWeight><ImageWeight>0<\ImageWeight>
<LinkWeight>0<LinkWeight>
```

Figure 1: Extracted structure from Figure 3



```
<Head>Support BCL Corpus <\Head> <ContentWeight>
<1412><\ContentWeight><ImageWeight>0<\ImageWeight><LinkWeight>0<LinkWeight>
```

Figure 2: Extracted structure from Figure 4

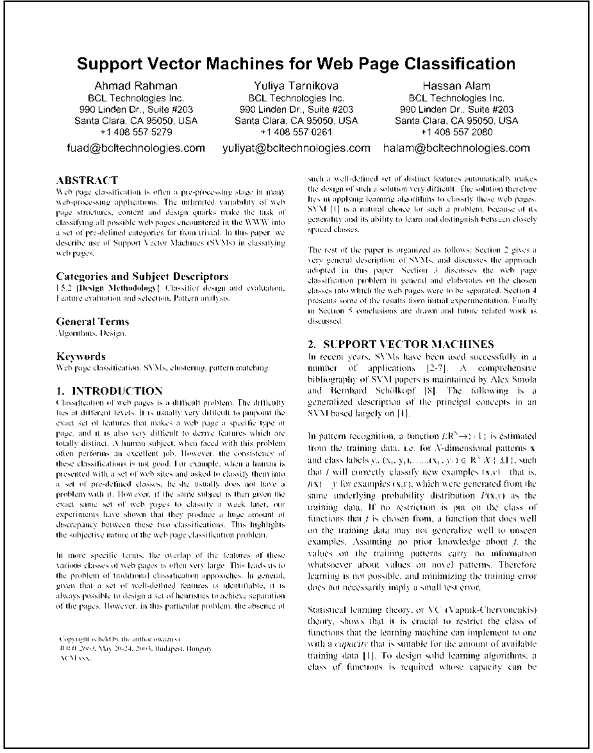
3.3 Creating a textual summary

Lexical chains have been used to create summaries of their content. Cohesion is a way of connecting different parts of text into a single theme. In other words, this is a list of semantically related words, constructed by the use of co-reference, ellipses and conjunctions. This aims to identify the relationship between words that tend to co-occur in the same lexical context. An example might be the relationship between the words "students" and "class" in the sentence: "The students are in class".

For every sentence in the node (the "content"), all nouns are extracted using a Parts of Speech (POS) tagger [14], all possible synonym sets are determined that each noun could be part of. For every synonym set, a lexical chain is created by utilizing a list of words related to these nouns by WordNet relations [15]. Once lexical chains are created, a score for each chain is calculated using the following scoring criterion:

Score = Chain Size * Homogeneity Index
where,
ChainSize = $\sum_{\text{all chain entries (ch(i)) in the text}}$ w(ch(i)); representing how large the chain is, and each member contributing according to how related it is.
w(ch(i)) = relation(ch(i)) / (1 + distance (ch(i)))
relation(ch(i)) = 1, if ch(i) is a synonym,
0.7, if ch(i) is an antonym,
0.4, if ch(i) is a hypernym, holonym or hyponym.
distance(ch(i)) = number of intermediate nodes in the hypernym graph for hypernyms and hyponyms and 0 otherwise.
Homogeneity Index = 1.5 – ($\sum_{\text{all distinct chain entries (ch(i)) in the text}}$ w(ch(i)))/ChainSize; representing how diverse the members of the chain are.

To make sure that there is no duplicate chain and that no two chains overlap, only one lexical chain with highest score is selected for every word and the rest are discarded.



Support Vector Machines for Web Page Classification

Ahmad Rahman
BCL Technologies Inc.
990 Linden Dr., Suite #203
Santa Clara, CA 95050, USA
+1 408 557 5279
fuad@bcltechnologies.com

Yuliya Tarnikova
BCL Technologies Inc.
990 Linden Dr., Suite #203
Santa Clara, CA 95050, USA
+1 408 557 0261
yuliyat@bcltechnologies.com

Hassan Alam
BCL Technologies Inc.
990 Linden Dr., Suite #203
Santa Clara, CA 95050, USA
+1 408 557 2080
halam@bcltechnologies.com

ABSTRACT

Web page classification is often a pre-processing stage in many web-processing applications. The unlimited variability of web page structures, content and design quarks make the task of classifying all possible web pages encountered in the WWW into a set of pre-defined categories far from trivial. In this paper, we describe use of Support Vector Machines (SVMs) in classifying web pages.

Categories and Subject Descriptors

I.5.2 [**Design Methodology**]: Classifier design and evaluation, Feature evaluation and selection, Pattern analysis.

General Terms

Algorithms, Design.

Keywords

Web page classification, SVMs, clustering, pattern matching.

1. INTRODUCTION

Classification of web pages is a difficult problem. The difficulty lies at different levels. It is usually very difficult to pinpoint the exact set of features that makes a web page a specific type of page, and it is also very difficult to derive features which are totally distinct. A human subject, when faced with this problem often performs an excellent job. However, the consistency of these classifications is not good. For example, when a human is presented with a set of web sites and asked to classify them into a set of pre-defined classes, he/she usually does not have a problem with it. However, if the same subject is then given the exact same set of web pages to classify a week later, our experiments have shown that they produce a huge amount of discrepancy between these two classifications. This highlights the subjective nature of the web page classification problem.

In more specific terms, the overlap of the features of these various classes of web pages is often very large. This leads us to the problem of traditional classification approaches. In general, given that a set of well-defined features is identifiable, it is always possible to design a set of heuristics to achieve separation of the pages. However, in this particular problem, the absence of such a well-defined set of distinct features automatically makes the design of such a solution very difficult. The solution therefore lies in applying learning algorithms to classify these web pages. SVM [1] is a natural choice for such a problem, because of its generality and its ability to learn and distinguish between closely spaced classes.

The rest of the paper is organized as follows: Section 2 gives a very general description of SVMs, and discusses the approach adopted in this paper. Section 3 discusses the web page classification problem in general and elaborates on the chosen classes into which the web pages were to be separated. Section 4 presents some of the results from initial experimentation. Finally in Section 5 conclusions are drawn and future related work is discussed.

2. SUPPORT VECTOR MACHINES

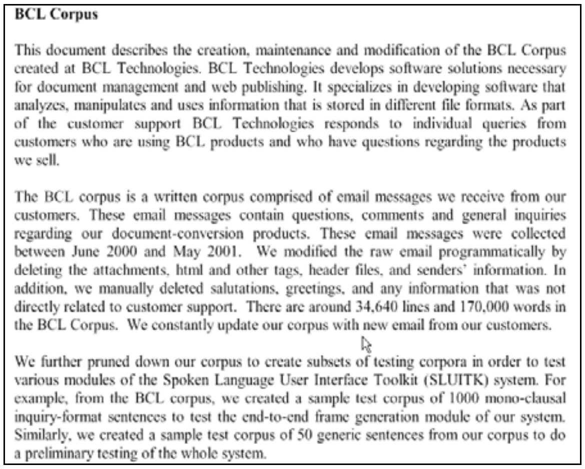
In recent years, SVMs have been used successfully in a number of applications [2-7]. A comprehensive bibliography of SVM papers is maintained by Alex Smola and Bernhard Schölkopf [8]. The following is a generalized description of the principal concepts in an SVM based largely on [1].

In pattern recognition, a function $f: R^N \rightarrow \{\pm 1\}$ is estimated from the training data, i.e. for N-dimensional patterns **x** and class labels y_i, $(x_1, y_1), \ldots, (x_l, y_l) \in R^N \times \{\pm 1\}$, such that f will correctly classify new examples (**x**,y) – that is, $f(\mathbf{x}) = y$ for examples (**x**,y), which were generated from the same underlying probability distribution $P(\mathbf{x},y)$ as the training data. If no restriction is put on the class of functions that f is chosen from, a function that does well on the training data may not generalize well to unseen examples. Assuming no prior knowledge about f, the values on the training patterns carry no information whatsoever about values on novel patterns. Therefore learning is not possible, and minimizing the training error does not necessarily imply a small test error.

Statistical learning theory, or VC (Vapnik-Chervonenkis) theory, shows that it is crucial to restrict the class of functions that the learning machine can implement to one with a *capacity* that is suitable for the amount of available training data [1]. To design solid learning algorithms, a class of functions is required whose capacity can be

Copyright is held by the author/owner(s).
WWW 2003, May 20-24, 2003, Budapest, Hungary.
ACM xxx.

Figure 3: Structured document



BCL Corpus

This document describes the creation, maintenance and modification of the BCL Corpus created at BCL Technologies. BCL Technologies develops software solutions necessary for document management and web publishing. It specializes in developing software that analyzes, manipulates and uses information that is stored in different file formats. As part of the customer support BCL Technologies responds to individual queries from customers who are using BCL products and who have questions regarding the products we sell.

The BCL corpus is a written corpus comprised of email messages we receive from our customers. These email messages contain questions, comments and general inquiries regarding our document-conversion products. These email messages were collected between June 2000 and May 2001. We modified the raw email programmatically by deleting the attachments, html and other tags, header files, and senders' information. In addition, we manually deleted salutations, greetings, and any information that was not directly related to customer support. There are around 34,640 lines and 170,000 words in the BCL Corpus. We constantly update our corpus with new email from our customers.

We further pruned down our corpus to create subsets of testing corpora in order to test various modules of the Spoken Language User Interface Toolkit (SLUITK) system. For example, from the BCL corpus, we created a sample test corpus of 1000 mono-clausal inquiry-format sentences to test the end-to-end frame generation module of our system. Similarly, we created a sample test corpus of 50 generic sentences from our corpus to do a preliminary testing of the whole system.

Figure 4: Unstructured "flat" document

Of the remaining chains, "strong chains" are determined by applying the following criterion:

Score >= Average Score + 0.5 * Standard Deviation

While generating the summary, each sentence with the strong chains is cumulatively added to form a summary until there is no sentence with a "strong" chain is left. Each sentence is scored by the following criterion:

$(\sum_{\text{all chains passing through this sentence, ch – an entry in the chain that is from the sentence}} w(ch)*Score + 2 * \sum_{\text{all chains starting in this sentence}} w(ch)*Score) / \text{sentence length}$

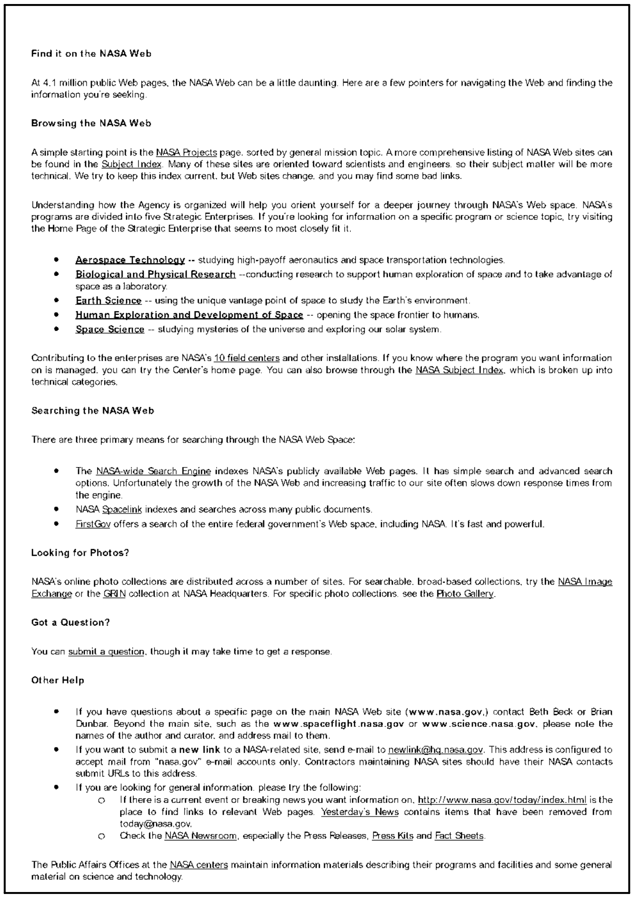
The final summary is formed by adding sentences to the summary starting with the highest score until there is no sentence left or the length or the summary reaches the target length. The target length of the summary is often related to the length of the original content, but can also be empirically set by the user.

3.4 Fusion of textual summary and the structure

Textual summaries can be combined in many ways with the structure of the document. This is primarily dictated by the requirement of the user, which in turn is governed by the way the summaries are to be used by the user. There are four different combination schemes:

- *Flat summary*: This is presented by combining the textual summary with only the title (if any) of the document. This is a quick way of converting a structured document to a non-structured summary form. This works best when the source document has a flat structure.
- *Distributed flat summary*: Sometimes the flat summary, when applied to structured documents, produces a skewed summary, i.e. some of the sections are heavily represented, but information from other sections is ignored. While this may be logical from the relevance of the content in terms of the overall content theme, the summary output often becomes hard to read. In distributed flat summary, each section is given its fair share of representation, calculated by associating the summary length of each section to the corresponding content weight. This is a quick way of converting a structured document to a flat summary form and works best when the source document is structured with uneven content distribution.
- *Structured summary*: This is presented by combining the textual summary with overall structure of the document. This preserves the structure of the original document and super-imposes the summary on that structure. This works best when the source document has a well-defined hierarchical structure, the content is evenly distributed and the composition is focused on a small number of themes.
- *Smart summary*: The summarizer automatically recommends the best possible type of summary and the optimum length by analyzing the document structure. For structured documents

with multiple levels of sections, it also recommends the number of levels in the summarizer output (e.g. 2nd level X.1, X.2 etc.).



Find it on the NASA Web

At 4.1 million public Web pages, the NASA Web can be a little daunting. Here are a few pointers for navigating the Web and finding the information you're seeking.

Browsing the NASA Web

A simple starting point is the NASA Projects page, sorted by general mission topic. A more comprehensive listing of NASA Web sites can be found in the Subject Index. Many of these sites are oriented toward scientists and engineers, so their subject matter will be more technical. We try to keep this index current, but Web sites change, and you may find some bad links.

Understanding how the Agency is organized will help you orient yourself for a deeper journey through NASA's Web space. NASA's programs are divided into five Strategic Enterprises. If you're looking for information on a specific program or science topic, try visiting the Home Page of the Strategic Enterprise that seems to most closely fit it.

- **Aerospace Technology** -- studying high-payoff aeronautics and space transportation technologies.
- **Biological and Physical Research** --conducting research to support human exploration of space and to take advantage of space as a laboratory.
- **Earth Science** -- using the unique vantage point of space to study the Earth's environment.
- **Human Exploration and Development of Space** -- opening the space frontier to humans.
- **Space Science** -- studying mysteries of the universe and exploring our solar system.

Contributing to the enterprises are NASA's 10 field centers and other installations. If you know where the program you want information on is managed, you can try the Center's home page. You can also browse through the NASA Subject Index, which is broken up into technical categories.

Searching the NASA Web

There are three primary means for searching through the NASA Web Space:

- The NASA-wide Search Engine indexes NASA's publicly available Web pages. It has simple search and advanced search options. Unfortunately the growth of the NASA Web and increasing traffic to our site often slows down response times from the engine.
- NASA Spacelink indexes and searches across many public documents.
- FirstGov offers a search of the entire federal government's Web space, including NASA. It's fast and powerful.

Looking for Photos?

NASA's online photo collections are distributed across a number of sites. For searchable, broad-based collections, try the NASA Image Exchange or the GRIN collection at NASA Headquarters. For specific photo collections, see the Photo Gallery.

Got a Question?

You can submit a question, though it may take time to get a response.

Other Help

- If you have questions about a specific page on the main NASA Web site (**www.nasa.gov**,) contact Beth Beck or Brian Dunbar. Beyond the main site, such as the **www.spaceflight.nasa.gov** or **www.science.nasa.gov**, please note the names of the author and curator, and address mail to them.
- If you want to submit a **new link** to a NASA-related site, send e-mail to newlink@hq.nasa.gov. This address is configured to accept mail from "nasa.gov" e-mail accounts only. Contractors maintaining NASA sites should have their NASA contacts submit URLs to this address.
- If you are looking for general information, please try the following:
  - If there is a current event or breaking news you want information on, http://www.nasa.gov/today/index.html is the place to find links to relevant Web pages. Yesterday's News contains items that have been removed from today@nasa.gov.
  - Check the NASA Newsroom, especially the Press Releases, Press Kits and Fact Sheets.

The Public Affairs Offices at the NASA centers maintain information materials describing their programs and facilities and some general material on science and technology.

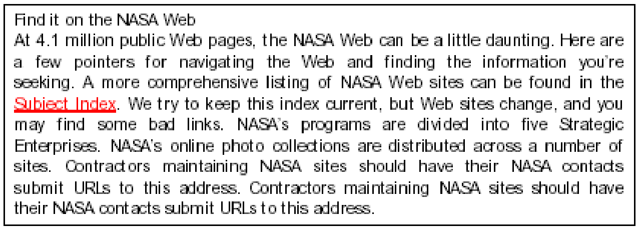
Figure 5: A sample document

4. Example

Figure 5 shows an example document. **Figure 6** shows a flat summary generated from this document. As expected, the document structure is lost, but the generated summary is coherent and meaningful. **Figure 7** shows a distributed flat summary of this document. This improves largely on the flat summary by exploiting information about the structure of the document. **Figure 8** shows the structured summary of the same document. This clearly shows that the summary retains the structure of the source document and that the summarization emphasizes the even distribution of the main theme. The readability of this summary is also the best of the three approaches for this particular example. The other types of summarization might be more appropriate based on the type of document.

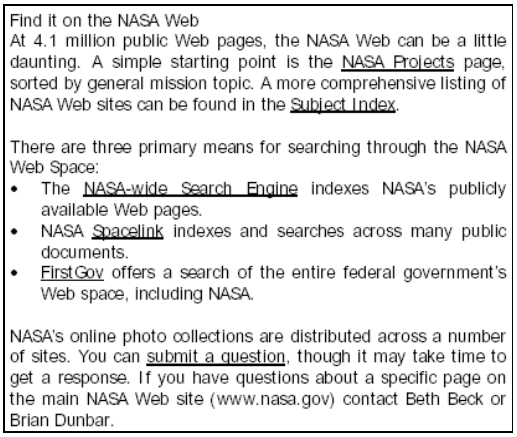
5. Evaluation

The proposed summarizer was compared with three summarizers commercially available, Copenic®, Sinope® and AutoSummarize®. A set of 22 documents were randomly collected by evaluators making sure that the samples included examples of both structured and flat documents. Evaluators assessed readability, ease of use, flexibility, customizability and accuracy of these summarizers. Overall, the proposed summarizer came out as either the top choice or the second choice in all of these categories. A sample set of their evaluation is presented in **Figure 9**.



Find it on the NASA Web
At 4.1 million public Web pages, the NASA Web can be a little daunting. Here are a few pointers for navigating the Web and finding the information you're seeking. A more comprehensive listing of NASA Web sites can be found in the Subject Index. We try to keep this index current, but Web sites change, and you may find some bad links. NASA's programs are divided into five Strategic Enterprises. NASA's online photo collections are distributed across a number of sites. Contractors maintaining NASA sites should have their NASA contacts submit URLs to this address. Contractors maintaining NASA sites should have their NASA contacts submit URLs to this address.

Figure 6: Flat summary



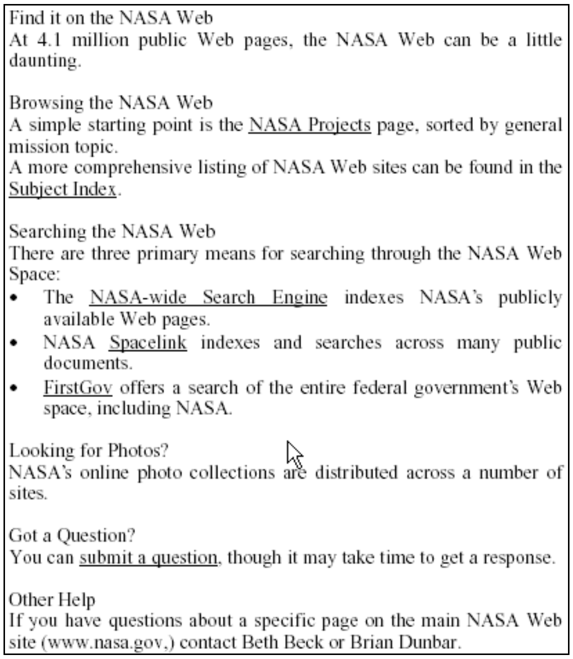
Find it on the NASA Web
At 4.1 million public Web pages, the NASA Web can be a little daunting. A simple starting point is the NASA Projects page, sorted by general mission topic. A more comprehensive listing of NASA Web sites can be found in the Subject Index.

There are three primary means for searching through the NASA Web Space:

- The NASA-wide Search Engine indexes NASA's publicly available Web pages.
- NASA Spacelink indexes and searches across many public documents.
- FirstGov offers a search of the entire federal government's Web space, including NASA.

NASA's online photo collections are distributed across a number of sites. You can submit a question, though it may take time to get a response. If you have questions about a specific page on the main NASA Web site (www.nasa.gov) contact Beth Beck or Brian Dunbar.

Figure 7: Distributed flat summary



Find it on the NASA Web
At 4.1 million public Web pages, the NASA Web can be a little daunting.

Browsing the NASA Web
A simple starting point is the NASA Projects page, sorted by general mission topic.
A more comprehensive listing of NASA Web sites can be found in the Subject Index.

Searching the NASA Web
There are three primary means for searching through the NASA Web Space:

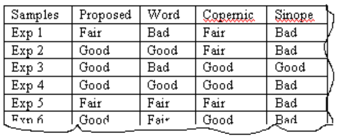
- The NASA-wide Search Engine indexes NASA's publicly available Web pages.
- NASA Spacelink indexes and searches across many public documents.
- FirstGov offers a search of the entire federal government's Web space, including NASA.

Looking for Photos?
NASA's online photo collections are distributed across a number of sites.

Got a Question?
You can submit a question, though it may take time to get a response.

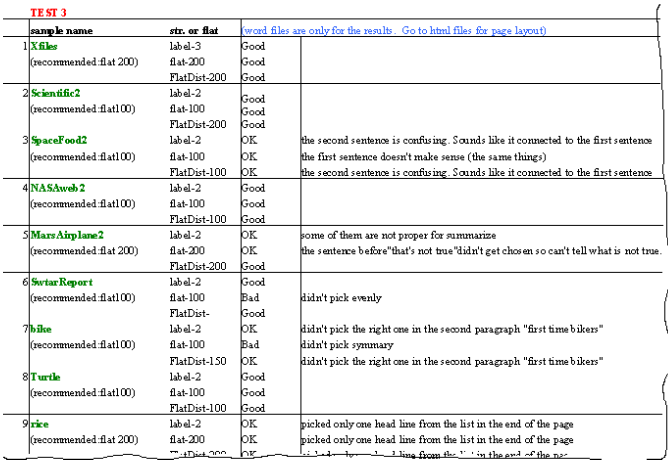
Other Help
If you have questions about a specific page on the main NASA Web site (www.nasa.gov,) contact Beth Beck or Brian Dunbar.

Figure 8: Structured summary



Samples	Proposed	Word	Copernic	Sinope
Exp 1	Fair	Bad	Fair	Bad
Exp 2	Good	Good	Fair	Bad
Exp 3	Good	Bad	Good	Good
Exp 4	Good	Good	Good	Bad
Exp 5	Fair	Fair	Fair	Bad
Exp 6	Good	Fair	Good	Bad

Figure 9: Part of the evaluation of the proposed summarizer to some commercial summarizer



TEST 3

	sample name	str. or flat	(word files are only for the results. Go to html files for page layout)	
1	Xfiles	label-3	Good	
	(recommended:flat 200)	flat-200	Good	
		FlatDist-200	Good	
2	Scientific2	label-2	Good	
	(recommended:flat100)	flat-100	Good	
		FlatDist-200	Good	
3	SpaceFood2	label-2	OK	the second sentence is confusing. Sounds like it connected to the first sentence
	(recommended:flat100)	flat-100	OK	the first sentence doesn't make sense (the same things)
		FlatDist-100	OK	the second sentence is confusing. Sounds like it connected to the first sentence
4	NASAweb2	label-2	Good	
	(recommended:flat100)	flat-100	Good	
		FlatDist-100	Good	
5	MarsAirplane2	label-2	OK	some of them are not proper for summarize
	(recommended:flat 200)	flat-200	OK	the sentence before"that's not true"didn't get chosen so can't tell what is not true.
		FlatDist-200	Good	
6	SwtarReport	label-2	Good	
	(recommended:flat100)	flat-100	Bad	didn't pick evenly
		FlatDist-	Good	
7	bike	label-2	OK	didn't pick the right one in the second paragraph "first time bikers"
	(recommended:flat100)	flat-100	Bad	didn't pick symmary
		FlatDist-150	OK	didn't pick the right one in the second paragraph "first time bikers"
8	Turtle	label-2	Good	
	(recommended:flat100)	flat-100	Good	
		FlatDist-100	Good	
9	rice	label-2	OK	picked only one head line from the list in the end of the page
	(recommended:flat 200)	flat-200	OK	picked only one head line from the list in the end of the page
		FlatDist-200	OK	picked only one head line from the list in the end of the pa

Figure 10: A part of the evaluation of the three types of summarization

The evaluators also compared the various options offered by the proposed summarizer on the same database. It was noted that flat documents were almost always better summarizer using the flat summary, but structured documents were better summarized by either the distributed flat summary or the structured summary, depending on how evenly the content is distributed and how many central themes were present in the document. **Figure 10** shows a snapshot of this evaluation.

6. Further work

The summarizer reported in this paper is a work in progress. Some of the issues that still need to be addressed include ways to generate a good summary from documents that have multiple main themes, have specific constructs such as bullets and lists, cross comparing section headings with text, and detecting relationship among sections for safe merging. Integration of this NLP summarization method to existing web page summarization techniques based on structural analysis alone [10,16] is already well underway [17].

6. Conclusion

This paper has presented a novel approach of summarizing structured and non-structured documents using a hybrid approach of structural analysis and theme generation using lexical chain computation. It was compared with three other summarizers commercially available and found to be either better or comparable to them.

References

[1] Knight, K. and Marcu, D. "Statistics-Based Summarization - Step One: Sentence Compression". AAAI/IAAI, pages 703-710, 2000.

[2] McKeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Kan, M., Schiffman, B., and Teufel, S. "Columbia Multi-Document Summarization: Approach and Evaluation". Workshop on Text Summarization, 2001.

[3] Brunn, M., Chali, Y., and Pinchak. C. "Text Summarization Using Lexical Chains". Work. on Text Summarization. 2001.

[4] Boguraev, B. and Neff, M. "Discourse Segmentation in Aid of Document Summarization". In Proceedings of Hawaii Int. Conf. on System Sciences (HICSS-33), Minitrack on Digital Documents Understanding, IEEE. 2000.

[5] Aho, A., Chang, S., McKeown, K., Radev, D., Smith, J., and Zaman, K. "Columbia Digital News Project: An Environment for Briefing and Search over Multimedia". Information J. - Int. J. on Digital Libraries, 1(4):377-385. 1997.

[6] Berger A. and Mittal, V. "Query-relevant summarization using FAQs". Proc. of the 38th Annual Meeting of the Association for Computational Linguistics. 2000.

[7] Barzilay, R., McKeown, K. and Elhadad, M. "Information fusion in the context of multi-document summarization". In Proc. of ACL'99, 1999.

[8] Yong Rui, Y., Gupta, A., and Acero, A. "Automatically extracting highlights for TV Baseball programs". ACM Multimedia, Pages 105-115, 2000.

[9] Shahar, Y. and Cheng, C. "Knowledge-based Visualization of Time Oriented Clinical Data". Proc AMIA Annual Fall Symp., pages 155-9, 1998.

[10] Rahman, A, H. Alam, R. Hartono and K. Ariyoshi. "Automatic Summarization of Web Content to Smaller Display Devices", 6th Int. Conf. on Document Analysis and Recognition, ICDAR01, pages 1064-1068, 2001.

[11] Harabagiu, S and Lacatusu, F. "Generating Single and Multi-Document Summaries with GISTexter". Document Understanding Conference 2002 (DUC 2002), 2002.

[12] Hovy, E. and Lin, C. "Automated text summarization in SUMMARIST". In Inderjeet Mani and Mark T. Maybury, editors, Advances in Automatic Text Summarization, chapter 8, pages 81-94. MIT Press. 1999.

[13] NIST web site on suumarization: http://www-nlpir.nist.gov/projects/duc/pubs.html, Columbia University Summarization Resources (http://www.cs.columbia.edu/~h-jing/summarization.html) and Okumura-Lab Resources (http://capella.kuee.kyoto-u.ac.jp/index_e.html).

[14] Brill E. "A Simple Rule-based Part of Speech Tagger". In Proc. of the 3rd Conference on Applied Natural Language Processing, 1992.

[15] WordNet - A lexical database for the English language. http://www.cogsci.princeton.edu/~wn/.

[16] Alam, H., Hartono, R. and Rahman, A. "Extraction and Management of Content from Html Documents". Chapter in the book tilted "Web Document Analysis: Challenges and Opportunities". World Scientific Series in Machine Perception and Artificial Intelligence, 2002. In press.

[17] Alam, H., Hartono. R., Kumar, A., Tarnikova, Y., Rahman, F. and Wilcox, C. "Web Page Summarization for Handheld Devices: A Natural Language Approach", 7th Int. Conf. on Document Analysis and Recognition (ICDAR'03).