# A FRAMEWORK FOR SUMMARIZATION OF MULTI-TOPIC WEB SITES

by

Yongzheng Zhang

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
August 2007

*Your file Votre référence*
*ISBN: 978-0-494-31513-2*
*Our file Notre référence*
*ISBN: 978-0-494-31513-2*

DALHOUSIE UNIVERSITY

To comply with the Canadian Privacy Act the National Library of Canada has requested that the following pages be removed from this copy of the thesis:

Preliminary Pages

Examiners Signature Page (pii)

Dalhousie Library Copyright Agreement (piii)

Appendices

Copyright Releases (if applicable)

# Table of Contents

## List of Tables

## List of Figures

# Abstract

Web site summarization, which identifies the essential content covered in a given Web site, plays an important role in Web information management. However, straightforward summarization of an entire Web site, which is large and with diverse content, may lead to a summary heavily biased to a subset of main topics covered in the target Web site. In this thesis, we propose a two-stage framework for effective summarization of multi-topic Web sites. The first stage identifies the main topics covered in a Web site and the second stage summarizes each topic separately.

In order to identify the different topics covered in a Web site, we perform both text- and link-based clustering. In text-based clustering, we investigate the impact of document representation and feature selection on the clustering quality. In link-based clustering, we study co-citation and bibliographic coupling. We demonstrate that text-based clustering based on the selection of features with high variance over Web pages is reliable and that outgoing links can be used to improve the clustering quality if a rich set of cross links is available.

Each individual cluster computed above is summarized using an extraction-based summarization system, which extracts key phrases and key sentences from source documents to generate a summary. The performance of such an extraction-based Web site summarization system depends on its underlying key phrase extraction method. Hence, we conduct a user study to investigate five alternative key phrase extraction methods. Results show that the best method combines linguistic constraints with frequency over the corpus adjusted to take into account nesting of terms. Another important component in an extraction based summarization system is the key sentence extraction. To this end, we design and develop a classification approach in the cluster summarization stage. The classifier uses statistical and linguistic features to determine the topical significance of each sentence.

Finally, we evaluate the proposed system via a user study. We demonstrate that the proposed clustering summarization approach significantly outperforms the single-topic summarization approach for any given Web site summarization task.

# List of Abbreviations and Symbols Used

| | |
|---|---|
| $A$ | Accuracy |
| $C$ | A cluster of documents |
| $E$ | Entropy |
| $G$ | Information gain of a phrase |
| $I$ | Mutual information of a phrase |
| $T$ | A set of documents with the same topic |
| $V$ | Term variance of a phrase |
| $W$ | Normalized weight of a phrase |
| $n$ | Document frequency of a phrase |
| $w$ | Weight of a phrase |
| **CHI** | $\chi^2$ Statistic |
| **CNC** | C-value/NC-value |
| **DF** | Document Frequency |
| **IG** | Information Gain |
| **KEA** | Automatic Keyphrase Extraction |
| **KWD** | Keyword |
| **MI** | Mutual Information |
| **MIX** | Mixture |
| **TFIDF** | Term Frequency Inverse Document Frequency |
| **TV** | Term Variance |

## Acknowledgements

This thesis could not have been completed without the help of many people. I appreciate all the contributions they have made for the thesis.

First of all, I would like to thank my co-supervisors, Dr. Evangelos Milios and Dr. Nur Zincir-Heywood, for their support and patience with me during the past five years. They have been the best friend and advisor a student could hope for. Their insight and inspiration for novel and practical research formed the foundation of this dissertation. I enjoyed all the motivating discussions to orientate my research towards the Web site clustering and summarization problem. Without their guidance, support, care, and patience, I could not have been where I am today.

I would also like to express my deepest appreciation to my committee members, Dr. Ali Ghorbani, Dr. Michael Shepherd and Dr. Jeannette Janssen, for their valuable and insightful comments on this research. Dr. Shepherd has played a unique and important role, providing countless comments and suggestions on my research. I am grateful to him for his help and feedback through the years.

I am thankful to Nawaaz Ahmed and Roy Shan at Yahoo!. Nawaaz was my mentor when I was a summer intern with the Yahoo! Search group in 2005. He provided great ideas and a stimulating environment during the summer. I enjoyed the opportunity to learn how big search engine works and to apply the summarization techniques to search relevance problems. Roy was my colleague and I am very grateful to him for all the motivating discussions on summarization problems. I was very fortunate to meet such great people outside of Dalhousie University.

Throughout my doctoral study, I benefited from numerous discussions with my friends and colleagues, who shared their wisdom and views from the perspective of their fields of expertise. I am grateful to all of them for their time and assistance. I owe my most direct thanks to my longtime friends Bin Tang and Lei Dong. They helped me on things too many to mention. Bin passed away in a tragic accident in 2005. I miss him greatly.

The years at Dalhousie would not have been nearly as much fun without the other

current and former members of the Machine Learning and Networked Information Spaces (MALNIS) Laboratory, the Dalhousie Natural Language Processing (DNLP) Group, the Web Information Filtering Lab (WIFL), and the Exploring Dynamic Groupware Environments (EDGE) Lab. Special thanks go to Dr. Kori Inkpen as well as my colleagues Kirstie Hawkey and Melanie Kellar for their valuable help on statistical tests. Furthermore, I also wish to thank the faculty and staff of the Faculty of Computer Science for their help and support during the last five years.

I am especially thankful to my wife Chen for her unconditional love, for putting up with the countless hours I spent on my thesis work, and for being there when I needed it. Frankly speaking, it is quite tough to be the wife of a graduate student, but she went through the last five years without many complaints. I also wish to thank my little girl Mia. She has been bringing us so much fun and happiness. Every time I see her smiling face, I feel so encouraged and energized to finish my thesis work.

Finally, I owe my foremost thanks to my parents, Xueying and Donghai, for their immeasurable love and support which enabled me to succeed. No words in any natural language would be sufficient to thank my parents for all they have done for me. This thesis is dedicated to them. Similarly, I would like to thank my entire family from my parents-in-law to my sister and brother-in-law for their love, encouragement and support in my Ph.D. journey.

# Chapter 1

# Introduction

In this chapter, we briefly describe the research problem, the motivation, and the approach. We aim to address the following two questions:

1. Why is Web site summarization important?

2. How to summarize a Web site with multiple topics?

## 1.1 Why Summarization

In recent years, the World Wide Web (WWW) has experienced a tremendous explosion of online information, which poses a great challenge for Web users to take full advantage of the huge information repository. Hence, effective management of online information becomes more and more critical.

Web information management involves the design of models for effective management of semi-structured data, meta-data, multimedia information, and multi-dimensional Web databases in order for Web users to precisely and quickly retrieve, locate, navigate, and visualize Web contents. Information management in the WWW context requires a set of tools, as shown in Figure 1.1. For example, Web Indexing and Retrieval is important for information seeking. Web users are often unable to obtain the information they want without the help of search engines (e.g. Google[1]), which have been gaining more popularity. Web document categorization usually provides a directory with hierarchical categories (e.g. Yahoo! Directory[2]) which helps Web users locate a particular kind of Web sites more quickly and effectively.

In this thesis, we focus on the summarization approach. Text summarization is the process of generating a concise yet meaningful summary which highlights the core contents of source documents. Web document summarization, which is derived from

[1] `http://www.google.com`
[2] `http://dir.yahoo.com`

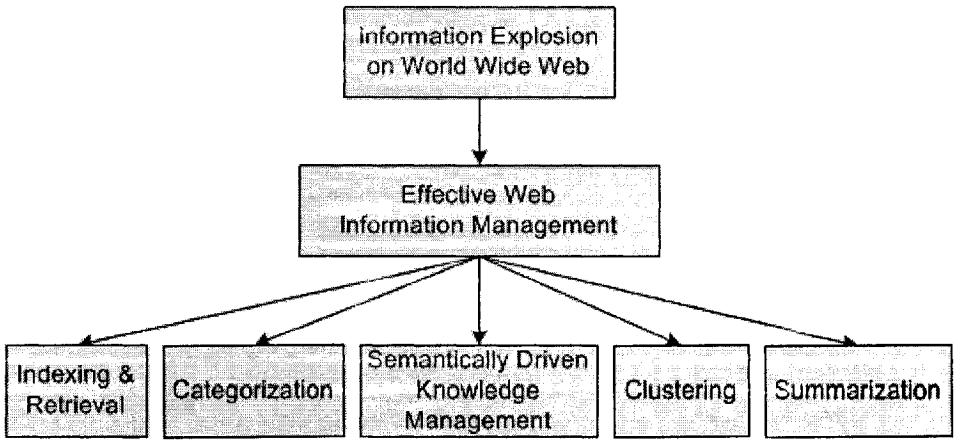


Figure 1.1: Typical approaches to information management on the World Wide Web.

traditional text summarization, is of particular interest and has played an important role in Web information management. It can be categorized as single Web document summarization and multiple Web document summarization.

### 1.1.1 Single Web Document Summarization

Single Web document summarization is the process of summarizing a given Web page by analyzing its textual content and/or contextual knowledge such as anchor text. It has the following main applications:

- Display on hand-held devices [8]. As the hand-held devices (e.g. PDA) become more and more popular, there is a rapidly increasing demand for display of Web pages on these devices. However, hand-held devices often have limited memory and computational power. Thus, display of conventional fancy pages is often a real problem. Also it is very expensive to design small Web pages that are specifically fit on these devices. It turns out that a short and meaningful summary of these conventional Web pages is a good alternative for display on hand-held devices.

- Query formation and expansion. When Web users come to search engines, they often have difficulty forming an accurate query to fulfill their information seeking goals. One popular reason is that they cannot find or think of all proper keywords that can be used to form a good query. If each result page of the initial query has a corresponding summary, which contains a few key phrases and key

sentences, then Web users will obtain clues from the summary to re-formulate or expand their initial queries. The new queries are often more accurate and can achieve better search results. The process can be repeated and it serves as a feedback system.

- Web document indexing. When search engines display the search results of a particular query, they present a snippet of each result document, which consists of a few sentences extracted from the document in which the query keywords appear. In fact, a Web document can be summarized during the crawling stage. The summary can then be stored as an alternative to the conventional snippet when the Web document is indexed. The concise summary will provide a more comprehensive overview of the document and help Web users more easily decide whether the document is what they are looking for.

- Document relevance ranking. Web search engines often return too many matching documents for a search query. Hence, an estimate of the relevance of the documents to the query should be provided such that more relevant documents show up near the top of the result list. A Web document summary can be used to improve the relevance ranking function. The query terms can be matched with the summary to obtain the popularity statistics (e.g. number of query terms that appear in the key phrase list and the key sentence list, positions of query terms in the summary). The obtained statistics can then be used to favor some Web documents over others.

### 1.1.2 Multiple Web Document Summarization

Multiple Web document summarization is an extension to the single Web document summarization task. It has been approached by using statistical and linguistic methods to create a summary which highlights the main contents covered in a Web document corpus. It has the following typical applications:

- Browsing of a Web site. A Web site summary can be used to browse the target site: a concise, informative, and meaningful summary can help Web users understand the essential topics and main contents covered in the Web site quickly without spending much browsing time [85].

- Organization of search engine results. Search engines often return hundreds of documents for a search query. It is difficult and time-consuming for Web users to browse even the top 10 result documents. It is much more user friendly to present a summary of the search results to the Web users.

- Organization of product reviews. As online shopping becomes more and more popular, shoppers tend to do more shopping research before making a purchase. Review sites (e.g. `http://reviews.cnet.com`), which either provide professional product reviews or allow Web users to write their own opinions and reviews, are gaining more and more traffic. It is very interesting to summarize all reviews of a particular product to show the main theme of these reviews, such as pros/cons, in addition to presenting all individual reviews.

- Web directory construction. DMOZ[3] and Yahoo! directories provide a hierarchy of categories, to which millions of Web sites are categorized. Each listed site has a concise human-written summary that highlights the essential content in order to help readers better understand the Web site. It is very expensive to manually author such a summary for each Web site. Alternatively, the concise summary can be obtained by summarizing a given site automatically.

## 1.2 How to Summarize

In this thesis, we focus on the Web site summarization task in the context of Web information management. Automatically generating coherent summaries as good as human-authored summaries is a challenging task since Web sites often contain diverse topics and heterogeneous contents. The size, diversity, and complexity of Web sites are continuing to grow at a fast rate.

In **single-topic summarization** (e.g. [85]), all Web pages in a given Web site are assumed to be in the same topic group (which is often inaccurate) and therefore are summarized directly. Such a **straightforward summarization** of the entire Web site often yields an incoherent summary or a summary that is heavily biased towards a subset of the topics included in the Web site.

[3] `http://www.dmoz.org`

Our main objective is to propose a system which can effectively summarize Web sites with multiple topics and heterogeneous contents to facilitate Web information management. In order to achieve this, we need the ability to first detect what the important topics are in a given Web site. It would be greatly helpful if we could detect the topical relationship between Web pages and group them accordingly. Site maps and index pages help a lot, but they do not always exist and are not always topically grouped.

We propose in this thesis a framework for effective summarization of multi-topic Web sites. The system first crawls a given Web site using the breadth-first search algorithm to build a link hierarchy. Each node in the link hierarchy represents a unique Web page collected in the Web site traversal. Then $K$-means or $X$-means clustering using coupled text- and link-based features is applied to identify the main topics included in the target Web site. Next, each individual cluster is separately summarized by our previous extraction-based summarization system [85]. Finally, the Web site summary consists of a few concise cluster summaries. The overview of this framework is shown in Figure 1.2.

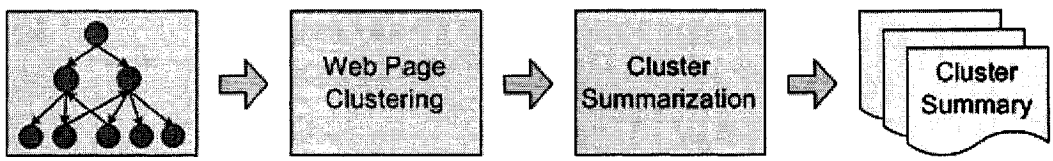


Figure 1.2: A framework for summarization of multi-topic Web sites.

The main objective of the Web Page Clustering component is to group documents into clusters where documents within the same cluster are related to each other whereas documents between different clusters are not related. We aim to investigate whether basic clustering algorithms such as $K$-means and $X$-means can find topic groups effectively. Critical evaluation of various clustering algorithms to find the best one in this task is a topic for future research.

Clustering approaches have been mostly text-based. Link analysis has been widely studied in many research areas such as Web document ranking [6, 35], Web document classification [23], topic distillation [10], document similarity analysis [46], Web structure mining [9], site map construction [41], and Web community identification

[18]. In this work, we aim to utilize both text- and link-based features. In text-based clustering, we investigate the impact of document representation and feature selection on the clustering quality. In link-based clustering, we employ co-citation and bibliographic coupling. We use entropy and accuracy to evaluate the clustering quality.

Summarization of an individual cluster is a multi-stage process following our previous single-topic summarization system [85]. The process involves five steps. First, plain text is extracted from the HTML source of Web pages. Second, text classification is performed to extract the narrative text[4] for more effective summarization. Third, key phrases are extracted from the narrative text in consideration. Fourth, key sentences are extracted from the narrative text based on the density of key phrases. Finally, a cluster summary is created consisting of both key phrases and key sentences. The cluster summarization process is shown in Figure 1.3.

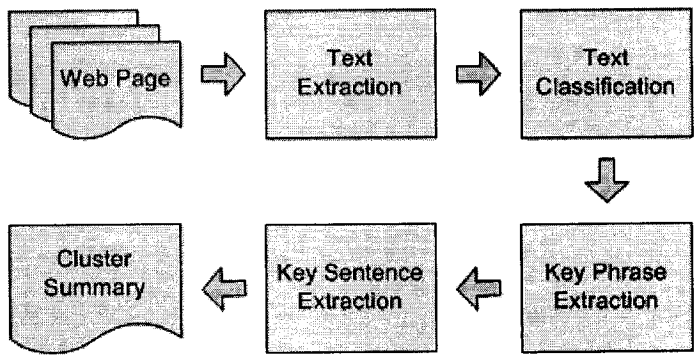


Figure 1.3: Typical approach to summarization of an Web document corpus.

The cluster summarization described above relies on the extraction of the most significant sentences from the target cluster based on the density of a list of key phrases that best describe the entire cluster. The performance of such an extraction-based approach heavily depends on its underlying key phrase extraction method. Therefore, it is critical to investigate alternative key phrase extraction methods in order to choose the best one. In this thesis, we benchmark five key phrase extraction

[4]Narrative text is the text paragraphs that are often more structured, informative and coherent than non-narrative text. Here is an example of a narrative paragraph: *The Software Engineering Process Group (SEPGSM) Conference is the leading international conference and exhibit showcase for software process improvement (SPI).* In contrast, a non-narrative paragraph often consists of short phrases or bullets, e.g. *First created on 10 May 2000. Last Modified on 22 July 2003. Copyright ©2000-2003 Software Archive Foundation. All rights reserved.*

methods that can be used in the key phrase extraction stage. The five methods are Term Frequency Inverse Document Frequency (TFIDF) [62], Automatic Keyphrase Extraction (KEA) [81], Keyword (KWD) [85], C-value/NC-value (CNC) [22], and Mixture (MIX), as shown in Figure 1.4.

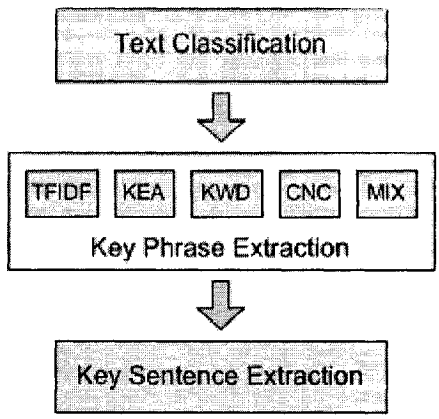


Figure 1.4: Five methods used in the key phrase extraction stage.

We aim to investigate how well each key phrase extraction method performs in the automatic Web site summarization task via a formal user study, i.e., we are interested in learning which method will yield summaries with the best quality. In our study, human subjects are asked to read the summaries generated based on different key phrase extraction methods. Then they rate each summary element using a 1-to-5 scale. The quality of each summary is calculated using both the *acceptable percentage* measure and the *quality value* measure. Acceptable percentage is the ratio of summary elements that receive a score of 3, 4 or 5. Quality value is the average score of a summary element. The quality of each type of summaries is compared with each other. One-Way Fully Repeated Measures ANOVA is used to conduct the statistical analysis.

Once the key phrases are identified for each cluster, key sentences can be retrieved from the narrative text. Traditionally, the key sentences are extracted based on the density of key phrases. In this thesis, we choose the best key phrase extraction method identified in the comparative study of key phrase extraction methods. Furthermore, we propose a classification approach to key sentence extraction in the cluster summarization stage. The classifier uses statistical and linguistic features to determine the topical significance of each sentence.

To investigate whether the clustering-summarization framework summarizes Web sites better than the single-topic summarization system, we also conduct a formal user study where human subjects are asked to read *the short cluster summaries* generated by our clustering-summarization framework, and *the single long summary* created by our previous summarization system based on the best key phrase extraction method. Then they rate each summary element using a 1-to-5 scale. Again, the quality of the short cluster summaries is calculated using both the acceptable percentage measure and the quality value measure.

The summarization framework proposed above consists of several key components, including feature selection and link analysis in Web page clustering, key phrase extraction and key sentence extraction in cluster summarization. More specifically, link analysis takes into account incoming and outgoing links when measuring the similarity of two Web documents. Key phrase extraction takes advantage of Web-based features in the following ways: 1) features in anchor text and special text are treated separately; 2) a KEA model for key term extraction is trained using Web documents; and 3) Web-specific stop words are used. Furthermore, key sentence classification uses the depth level of web pages in a web site. The contribution of this thesis is in aggregating these individual components and applying them to the Web site summarization problem.

The rest of the thesis is organized as follows: Chapter 2 reviews literature on Web document summarization. In Chapter 3, we discuss the comparative study of key phrase extraction methods in the Web site summarization task. Chapter 4 details how to perform coupled text- and link-based clustering. Then Chapter 5 describes the summarization of resulting clusters and presents the evaluation results. Finally Chapter 6 concludes our work and describes future work.

# Chapter 2

# Literature Review

This chapter reviews literature in text summarization and Web document summarization.

## 2.1 Automatic Text Summarization

Research in automatic text summarization dates back at least to 1958, when Luhn [47] proposed a simple approach which extracts significant sentences to form a summary based on features such as average term frequency and sentence location. Existing text summarization systems generate summaries automatically either by *extraction* or *abstraction.*

The goal of abstraction [3] is to understand the text using knowledge-based methods and compose a coherent summary comparable to a human authored summary. This is very difficult to achieve with current natural language processing techniques [24].

An easier alternative, extraction, has been the focus of automatic text summarization research in recent years [27, 31]. Extraction-based systems [11, 16, 25, 73] analyze source documents using techniques derived from information retrieval (e.g. frequency analysis and keyword identification) to determine the most significant sentences that constitute the summary. The significance of a sentence is determined by features such as the density of keywords [85] and rhetorical relations [51] in the context.

Kupiec et al. [38] build a statistical classification model based on training documents with hand-selected extracts. The model estimates the probability that a given sentence is included in an extract based on a set of heuristic features. Summary generation for new documents proceeds with ranking sentences according to this probability.

Chuang and Yang [11] propose an approach which generates a summary automatically. First, sentences are broken into segments by special cue phrases. Next, each segment is represented by using a set of pre-defined features, both unstructured (e.g. title words) and structured (e.g. rhetorical relations). Finally, machine learning algorithms are applied to the feature set to extract the most important sentence segments for summary inclusion.

Post-processing of the extracted sentences has been used to produce succinct summaries without redundant information. For example, clustering has been applied to find clusters of closely related sentences, and only "core" sentences from all the clusters are used to form the output summary [29]. As a second example, the identification of named entities can help the system rephrase the pronouns used in order to create a meaningful summary [52].

Evaluation of automatically generated summaries can proceed in either of two different modes, *intrinsic* and *extrinsic*. Intrinsic evaluation compares automatically generated summaries against a gold standard (ideal summaries), which is very hard to construct. Extrinsic evaluation measures the utility of automatically generated summaries in performing a particular task (e.g. classification) [49, 69]. Extrinsic evaluation is also called task-based evaluation and it has become increasingly popular recently [60].

### 2.1.1 Multi-document Summarization

Multi-document summarization (MDS) is an extension of single-document summarization into a collection of documents [48]. Multi-document summaries can save users significant time in reading relevant text documents or browsing Web sites. Many of the single-document summarization techniques can also be used in multi-document summarization. However, issues such as *anti-redundancy* and *cohesion and coherence* become critical in MDS [25, 42]. Moreover, multi-document summarization lacks standard procedures and methodologies for evaluation, in contrast to the single-document summarization task [64].

The National Institute of Standards and Technology (NIST) sponsored the Document Understanding Conference[1] starting in 2001, which aims towards providing

[1] http://duc.nist.gov

standard training and test document collections (mostly news articles) which can be shared among the research community, as well as evaluations in single- and multi-document summarization for the conference participants [42].

Current MDS systems often apply a two-phase process, i.e., *topic identification* and *summary generation.* In the first phase, main topics (or events) covered in the multiple source documents are identified. Documents regarding the same topic (or event) with variations in presentation are put into the same set. Then each set of closely related documents is used to produce representative passages for the final summary by extraction or by reformulation [29, 71].

Radev et al. [60] present a MDS system called MEAD, which first uses modified TF-IDF measure to form clusters of documents on the same topic, and then uses centroids of the clusters to identify which sentences are most likely to be relevant to the cluster topic, rather than individual articles. Evaluation demonstrates that summaries generated by MEAD are as good as human created summaries.

Stein et al. [71] propose a different approach which first summarizes single documents and groups summaries in clusters, then selects representative passages from clusters, and finally organizes passages into a coherent summary.

McKeown et al. [52] introduce a system which first identifies the type of document sets, i.e. single-event, person-centered (or biographical), or multi-event, and then summarizes them accordingly.

One research area that is closely related to MDS is called topic hierarchy construction, where a hierarchy of topics represented by key terms is constructed. Sanderson and Croft [63] use a term association method to build a term hierarchy for a set of retrieved documents, with an ordering from general terms to more specific, i.e., the parent concept subsumes the child concept. Lawrie et al. [39] apply the Dominating Set algorithm to present a probabilistic language model for automatic topic hierarchy construction. Terms are efficiently chosen from a retrieved set and form a hierarchy serving as a multi-document summary. Lawrie and Croft [40] build statistical language models to recursively identify the most topical and predicative terms for hierarchy creation. Documents are attached to the hierarchy if they include topic terms. However, these approaches only work well on a small set of related documents. As for summarization, they do not provide key sentences, which is a normal experience

in multi-document summarization.

## 2.2 Web Document Summarization

Web document summarization is derived from traditional plain text summarization techniques [85]. To the best of our knowledge, research in Web document summarization has been primarily focused on summarization of a single Web page.

### 2.2.1 Web Page Summarization

Web page summarization has been either *context-based* or *content-based*. Context-based systems [2, 14] analyze and summarize the context of a Web document (e.g. brief content descriptions from search engine results) instead of its contents. Content-based systems [3, 8] derive from traditional text summarization techniques. The great challenge in Web page summarization is the diversity of contents and the frequent lack of a well-defined discourse structure compared to traditional text [3]. Approaches based on implicit document association (rhetorical relation) analysis [51] are difficult to apply to Web page summarization.

Amitay and Paris [2] propose an approach, which relies on the hypertext structure and the way information is described using it. Instead of analyzing the Web page itself, this approach collects the context of the document by tracing back-links, a service offered by search engines like Google. Text units which contain the link to the target Web page are then extracted. Finally, an automatic filter is used to select the best description for the Web page. Single-sentence sized coherent textual snippets are generated and presented to the user together with results from search engines Google and AltaVista[2]. The experiments show that on average users prefer the system to search engines.

Delort et al. [14] address three important issues, *contextualization*, *partiality*, and *topicality* faced by any context-based summarizer and propose two algorithms whose efficiency depends on the size of the text contents and the context of the target Web page.

The drawback of the systems that rely on context analysis is that context information of target pages is not always available and accessible. Consequently, approaches

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

which analyze source contents have been gaining more popularity. However, they rely on the underlying key phrase extraction method to generate key phrases in order to further identify key sentences.

Berger and Mittal [3] propose a system called OCELOT, which applies standard statistical models (in particular, the Expectation Maximization (EM) algorithm) to select and order words into a "gist", which serves as the summary of a Web document.

Buyukkokten et al. [8] compare alternative methods for summarizing Web pages for display on handheld devices. The *Keyword* method extracts keywords from the text units, and the *Summary* method identifies the most significant sentence of each text unit as a summary for the unit. They test the performance of these methods by asking human subjects to perform specific tasks using each method, and conclude that the combined *Keyword/Summary* method provides the best performance in terms of access times and number of pen actions on the hand held devices.

### 2.2.2 Web Site Summarization

In our previous work [85], we extend single Web document summarization to the summarization of complete Web sites. The "Keyword/Summary" idea of [8] is adopted, and the methodology is substantially enhanced and extended to Web sites as follows:

1. **Web Page URL Crawling** In order to summarize a given Web site, a certain number of Web pages within a short distance from the root (home page) of the target site, which are assumed to describe the main contents of the site in general terms, are collected by a specific Web crawler via the breadth-first search starting at the home page.

2. **Plain Text Extraction** After the Web pages have been collected, plain text is extracted from these Web pages and segmented into text paragraphs by the text browser *Lynx*[3], which is found to outperform several alternative text extraction tools such as *HTML2TXT*[4] and *html2txt*[5], in terms of more effective selection of plain text.

[3] `http://lynx.isc.org`
[4] `http://user.tninet.se/~jyc891w/software/html2txt`
[5] `http://cgi.w3.org/cgi-bin/html2txt`

3. **Narrative Text Classification** Since Web documents are often not well-structured with diverse contents such as tables of contents and link lists, it is important to determine which text paragraphs should be considered for summarization. This is achieved in two steps. First, a C5.0[6] classifier *LONGSHORT* is used to filter out *short* text paragraphs. Second, *long* paragraphs are classified into *narrative* or *non-narrative* by another C5.0 classifier *NARRATIVE*, and only narrative paragraphs are used in summary generation. These two classifiers are built based on features (e.g. number of words and part of speech tag) extracted by shallow natural language processing. The cross-validation shows a mean error of 5.9% and 11.3% for LONGSHORT and NARRATIVE, respectively.

4. **Key Phrase Extraction** Traditionally, key phrases (single keywords or multi-word keyterms) for the entire document corpus are extracted in order to generate a summary. Based on such key phrases, the most significant sentences, which best describe the source documents, can be retrieved. Key phrase extraction from a body of text relies on an evaluation of the importance of each candidate key phrase [8].

5. **Key Sentence Extraction** Once the key phrases are identified, the most significant sentences for summary generation are retrieved from all narrative paragraphs based on the presence density of key phrases [11].

6. **Summary Formation** The overall summary is formed by the top 25 key phrases and the top 5 key sentences. These numbers are empirically determined based on the fact that key sentences are more informative than key phrases, and the whole summary should fit in a single page.

[6] `http://www.rulequest.com/see5-unix.html`

# Chapter 3

# Comparing Key Phrase Extraction Methods

In this chapter, we describe our user study of comparing five key phrase extraction methods which can be used in our previous Web site summarization system [85].

## 3.1 Introduction

Up to date approaches to Web document summarization have often been extraction-based [3, 8, 85]. They apply statistical and linguistic analysis to extract key phrases[1] which best describe the source documents, and further extract the most significant sentences based on the presence density of key phrases [8, 85].

In our previous work [85], we extend single Web document summarization to the summarization of complete Web sites. This approach generates a Web site summary consisting of the top 25 keywords and the top 5 key sentences. The system applies machine learning and shallow natural language processing techniques to extract the narrative text, and then extracts keywords from the narrative text together with anchor text and special text (e.g. emphasized text). The key sentences are identified based on the density of keywords. The evaluation shows that the automatically generated summaries are as informative as human authored summaries (e.g. DMOZ summaries).

The performance of an extraction-based Web site summarization system is mainly determined by its underlying key phrase extraction method. Automatic key phrase extraction has been a useful tool in many text related applications such as text clustering and document similarity analysis [54]. Traditional approaches to key phrase extraction are focused on frequency analysis such as TFIDF and collocation detection based on mutual information [50].

Recently, more effective systems have been developed. Krulwich and Burkey use

[1] A phrase can be either a single word or a multi-word term. Throughout the thesis, we use keywords, keyterms, and key phrases interchangeably, depending on the method context.

heuristic rules such as the use of acronyms and the use of italics to extract key phrases from a document for use as features of automatic document classification [37]. Turney proposes GenEx, a key phrase extraction system, which consists of a set of parameterized heuristic rules that are tuned to the training documents by a genetic program [75]. However, these methods heavily depend on heuristic rule pre-defining and tuning.

In this work, we investigate five key phrase extraction methods, i.e., Term Frequency Inverse Document Frequency (TFIDF) [62], Automatic Keyphrase Extraction (KEA) [81], Keyword (KWD) [85], C-value/NC-value (CNC) [22], and Mixture (MIX). These methods have been well studied in related literature [21, 26, 32, 33, 46, 54, 57, 61, 76, 77, 83, 84, 86] and can be used in the key phrase extraction stage of our Web site summarization system [85] described above.

- The TFIDF method captures a word's frequency in a single document compared to its rarity in the whole document collection. It has been widely studied in many information retrieval tasks so we use it as the baseline method.

- The second method, KEA, builds a Naïve Bayes learning model using training documents with known key phrases, and then uses the model to find key phrases in new documents.

  We acknowledge that both TFIDF and KEA were originally designed for key phrase extraction from single documents so we extend them to the application on an entire document collection.

- The KWD method constructs a C5.0 classifier using Web pages with known single keywords, and then uses this model to identify keywords from a new Web site.

- The fourth method, CNC, consists of both linguistic and statistical analysis to extract multi-word keyterms automatically.

  Both KWD and CNC are designed for key phrase extraction from an entire document collection.

- Finally, the MIX method combines KWD and CNC to obtain a list of key phrases.

Research in [74, 81] evaluates key phrase extraction methods by matching automatically extracted key phrases with human authored ones. In [77], Turney defines *acceptable* key phrases as good and fair key phrases, which are rated by human subjects. In the WWW context, manually identifying key phrases is time-consuming because of the diversity and complexity nature of Web documents.

In this thesis, we aim to investigate how well each key phrase extraction method performs in the automatic Web site summarization task. We conduct a user study where human subjects rate each summary element using a 1-to-5 scale. We are interested in learning which method yields summaries with the best quality.

We compare the key phrases generated by different methods in terms of both *acceptable percentage* and *quality value*. Acceptable percentage is the ratio of key phrases that receive a score of 3, 4 or 5. Quality value is the average score of a summary element. The quality of each type of summaries is compared with each other. One-Way Fully Repeated Measures ANOVA is used to conduct the statistical analysis.

## 3.2 Key Phrase Extraction

In this section, we explain in detail the five key phrase extraction methods, i.e., TFIDF, KEA, KWD, CNC, and MIX. These methods generate single keywords or multi-word keyterms or a mixture of the above two by a critical evaluation of the significance of each candidate key phrase in the source documents.

We realize that these methods have been designed to extract key phrases from traditional well-structured text such as technical papers and news articles. Research in [86] demonstrates that application of key phrase extraction on Web documents relies on the identification of narrative text, which often contains more structured, informative and coherent information than non-narrative text. Thus, we apply the NARRATIVE classifier introduced in [85] to extract narrative text. Then each key phrase extraction method will work on the narrative text only instead of all plain text.

### 3.2.1 TFIDF Method

TFIDF is a standard keyword identification method in information retrieval tasks (e.g. [20, 54, 66, 86]). It gives preference to words that have high frequency of occurrence in a single document but rarely appear in the whole document collection. In this work, we aim to use TFIDF as a baseline method to extract keywords from pages of a given Web site. This involves in the following steps:

1. For each Web page of the target Web site, identify the narrative text and convert it to lower case.

2. Extract all tokens in the narrative text, i.e., identify single words by removing punctuation marks and numbers. A standard set of 425 stop words (*a, about, above, ...*) [19] are discarded at this stage.

3. Apply Porter stemming to obtain word stems and update the number of documents in which each word stem appears.

4. Once all Web pages are processed using the above three steps, calculate the TFIDF value $w_{i,j}$ of word stem $i$ in page $j$ using the following equation:

$$w_{i,j} = \frac{n_{i,j}}{|p_j|} \cdot \log \frac{N}{n_i} \tag{3.1}$$

   where $\frac{n_{i,j}}{|p_j|}$ is the normalized term frequency of word stem $i$ in page $j$, $n_i$ is the number of pages that contain word stem $i$, and $N$ is the total number of Web pages in consideration.

5. For each Web page $j$, TFIDF value $W_{i,j}$ of each word stem in this page is normalized to unit length as follows:

$$W_{i,j} = \frac{w_{i,j}}{\sqrt{\sum_i {w_{i,j}}^2}}. \tag{3.2}$$

6. Finally, choose the top five word stems ranked by $W_{i,j}$ for each page. The number 5 is chosen based on the fact that often 3 to 5 key phrases are included in a technical article. Then replace each word stem with its original form which appears most frequently in the collection (e.g. "engin" ("engineering": 8, "engineer": 2) $\rightarrow$ engineering).

### Application of TFIDF on a Web Site

TFIDF is aimed towards extracting keywords from individual documents in a document collection rather than from the whole collection. Hence in order to generate a keyword list for an entire Web site, the output keywords from all pages should be combined properly. We aim to do the following:

1. Unite the 5 keywords from each Web page to obtain a single list. Each keyword (more precisely, its stem) $i$ has a normalized weight $W_{i,j}$, as shown in Equation 3.2.

2. Record $f_i$, which is the number of pages in which keyword $i$ appears[2]. Let $W_i$ be the overall weight of keyword $i$ in the Web site and $A_i$ be its average weight. So $W_i = \sum_j W_{i,j}$, and $A_i = W_i / f_i$.

3. Now three features, i.e., $W_i$, $A_i$, and $f_i$ can be used to re-rank the list in order to select the top 25 keywords for the target Web site. The number 25 is an empirical number set in the summarization framework [85]. Preliminary tests show that in terms of acceptable percentage (see Paragraph 3.3.3), $f_i$ is the best feature.

4. The top 25 keywords are taken as the keywords for the target Web site and their weights are re-normalized for the purpose of key sentence extraction.

### 3.2.2 KEA Method

KEA [81] is an efficient and practical algorithm for extracting key phrases, i.e., single keywords and multi-word keyterms. It consists of two stages: "training" and "extraction". In the training stage, KEA builds a Naïve Bayes classifier using training documents with human-authored key phrases. More explicitly, KEA chooses a set of candidate key phrases from input documents. For each candidate, two feature values, $tfidf$ and *first occurrence*, are calculated. First occurrence is calculated as the number of words that precede the candidate's first appearance, divided by the

[2]This is again a document frequency concept. However, only those documents in which word stem $i$ serves as a keyword are counted.

total number of words in the document. This is the normalized distance to the beginning of the document. Those candidates that are human-authored key phrases are positive examples in the KEA model construction. In the extraction stage, KEA uses the classifier to find the best set of (by default 5) key phrases in new documents. More explicitly, KEA chooses a set of candidate key phrases from new documents and calculate the two feature values as above. Then each candidate is assigned a weight, which is the overall probability that this candidate is a key phrase.

### KEA Training

KEA is originally designed for key phrase extraction from traditional coherent text such as technical reports. In order to obtain a good KEA model for key phrase extraction from Web documents, we need to investigate whether KEA works well on diverse Web documents instead of traditional coherent text. Hence we build two KEA models as follows.

- The training set bundled with the Java-based KEA package (Version 3.0)[3] is used to train a *CSTR* KEA learning model. This data set contains 80 abstracts of Computer Science Technical Reports (CSTR) from the New Zealand Digital Library project[4]. Each abstract has 5 human-authored key phrases. The input to the Java program consists of text files with the corresponding key phrases. Research in [81] shows that a training set of 25 or more documents can achieve good performance.

- A total of 80 Web pages are randomly collected from 60 DMOZ Web sites. The criterion is that the Web page must have at least one narrative paragraph identified by the NARRATIVE classifier described in [85]. We browse each Web page and extract up to five key phrases from its narrative text. Then a *NTXT* (Narrative TeXT) KEA model is constructed.

Web pages are different from technical reports in terms of the diversity of contents and discourse structure. Hence, we intentionally choose technological Web pages in

[3] `http://www.nzdl.org/Kea`
[4] `http://www.nzdl.org`

order to eliminate the potential bias that the technical reports could have on building the CSTR model and to make these two models more comparable to each other.

We apply separately the CSTR model and the NTXT model to extract key phrases from the narrative text of Web pages. Preliminary experiments show that the NTXT model can extract key phrases with higher acceptable percentage so we use this model for key phrase extraction from Web pages of a given Web site.

#### Application of KEA on a Web Site

For the same reason as the application of TFIDF on an entire Web site, we aim to do the following:

1. Unite the 5 key phrases from each Web page to obtain a single list. Each key phrase $i$ has a normalized weight $w_{i,j}$ in page $j$, which is the overall probability value provided by the KEA model.

2. Compute the same three features, i.e., $W_i$, $A_i$, and $f_i$, as in the application of TFIDF. Preliminary tests again show that $f_i$ is the best feature in terms of acceptable percentage.

3. The top 25 phrases are chosen as the key phrases for the target Web site and their weights are re-normalized.

### 3.2.3 KWD Method

The KWD method introduced in [85] consists of two stages: C5.0 learning model construction and keyword identification. In both stages a set of candidate keywords are chosen from the target Web site, and then the values of certain features (e.g. frequency, part-of-speech tag) for each candidate keyword are calculated.

#### Learning Keywords

As discussed before, Web pages are different from traditional plain text documents. The existence of *anchor text* and *special text* (e.g. titles, headings, italic text) contributes much to the difference. Anchor text is the text associated with hyperlinks, and is often considered to be an accurate description of the Web page linked to. A

supervised learning approach is applied to learn the significance of each category of candidate keywords.

In order to produce decision tree rules for determining the keywords of given Web site, a data set of 5454 candidate keywords from the 60 DMOZ Web sites is collected. For each Web site, the frequency of each unique word (after stemming) in narrative text, anchor text and special text, is measured. Then the total frequency of each word over these three categories is computed, where the weight for each category is the same. Stop words are discarded at this stage.

For each candidate keyword, eight features of its frequency statistics (e.g. ratio of frequency to sum of frequency, ratio of frequency to maximum frequency in anchor text) in three text categories and the part-of-speech tag [5] are extracted. In particular, the weight of a candidate keyword is defined as *the ratio of its frequency (over three categories of text) to the sum of frequency of all candidate keywords.*

Next, each candidate keyword is labelled manually as *keyword* or *non-keyword.* The criterion to determine whether a candidate keyword is a true keyword is that the candidate must provide important information about the Web site. Based on frequency statistics and part-of-speech tags of these candidate keywords, a C5.0 classifier *KEYWORD* is constructed.

The resulting decision tree shows that anchor text and special text do play an important role in determining keywords of a Web site [85]. Among the total 5454 cases, 222 cases are misclassified, leading to an error of 4.1%. The ten-fold cross-validation of the classifier shows a mean error of 4.9%, which indicates the accuracy of this classifier.

### Keyword Identification

Once the decision tree rules for determining keywords have been built, they can be used to automatically extract keywords from a new Web site. First a list of candidate keywords is selected based on the same frequency analysis shown above and ranked by the weight. Then the classifier KEYWORD identifies all keywords in the list and the top 25 keywords are kept and used for key sentence extraction. It is observed that 40% to 70% of keywords appear in the home page of a Web site.

### 3.2.4 CNC Method

The KWD method is based on word frequency analysis against three different categories of text, i.e., narrative text, anchor text, and special text. This method is unable to extract terms consisting of multiple words. Since multi-word terms are more informative than single words [54], we aim to apply a state-of-the-art method C-value/NC-value [22] to extract multi-word keyterms from a Web site automatically, and further identify key sentences for summary generation.

#### Automatic Term Extraction

The CNC method consists of both linguistic analysis (linguistic filter, part-of-speech tagging [5], and stop-list) and statistical analysis (frequency analysis, *C-value*, *NC-value*) to extract and rank a list of terms by *NC-value*. A linguistic filter is used to extract word sequences likely to be terms, such as noun phrases and adjective phrases.

The C-value is a domain-independent method used to automatically extract multi-word terms from the whole document corpus. It aims to get more accurate terms than those obtained by the pure frequency of occurrence method, especially terms that may appear as nested within longer terms. *C-value* is formally represented in Equation 3.3.

$$Cv(a) = \begin{cases} \log_2 |a| f(a), \ a \text{ is not nested.} \\ \log_2 |a| (f(a) - \frac{\sum_{b \in T_a} f(b)}{P(T_a)}), \text{ otherwise.} \end{cases} \tag{3.3}$$

where $a$ is a candidate term; $|a|$ is the number of words in $a$; $f(a)$ is the frequency of occurrence of $a$ in the corpus; $T_a$ is the set of extracted candidate terms that contain $a$; and $P(T_a)$ is the number of these longer candidate terms.

The NC-value is an extension to the C-value, which incorporates information of context words into term extraction. Context words are those that appear in the vicinity of candidate terms, i.e. nouns, verbs and adjectives that either precede or follow the candidate term. Each context word is assigned a weight as follows:

$$weight(w) = \frac{t(w)}{n} \tag{3.4}$$

where, $w$ is a term context word (noun, verb or adjective); $weight(w)$ is the assigned weight to the word $w$; $t(w)$ is the number of terms the word $w$ appears with; and $n$ is

the total number of terms considered and it expresses the weight as the probability that the word $w$ might be a term context word.

*NC-value* is formally given by Equation 3.5.

$$NCv(a) = 0.8 \times Cv(a) + 0.2 \times \sum_{b \in C_a} f_a(b) \cdot weight(b) \tag{3.5}$$

where $a$ is a candidate term; $C_a$ is the set of distinct context words of $a$; $b$ is a word from $C_a$; $f_a(b)$ is the frequency of $b$ as a term context word of $a$; and $weight(b)$ is the weight of $b$ as a term context word. The two components of the *NC-value*, i.e., *C-value* and the context information factor, have been assigned the weights 0.8 and 0.2, respectively. These two coefficients were derived empirically [22].

Experiments in [22, 54] show that the CNC method performs well on a variety of special text corpora. In particular, with the open linguistic filter `(Adj.|Noun)+Noun` (one or more adjectives or nouns followed by one noun), the CNC method extracts more terms than with the closed linguistic filter `Noun+Noun` (one or more nouns followed by a noun) without much precision loss. For example, terms such as *artificial intelligence* and *natural language processing* will be extracted by the open linguistic filter. Hence, in our work, we use this linguistic filter to extract terms from a Web site.

#### Keyterm Identification

The candidate term list $C$ (ranked by *NC-value*) of a Web site contains some noun phrases (e.g. *privacy statement*), which, although they appear frequently in various Web sites, are not relevant to the core content of the Web sites and hence must be treated as Web-specific stop words [67]. We experimented with the 60 DMOZ Web sites used in the KWD method and manually identified a stop list, $L$, of 51 noun phrases (e.g. *Web site*) [83]. The candidate term list $C$ is filtered through the noun phrase stop list $L$, and only the top 25 terms (ranked by *NC-value*) are selected as keyterms.

### 3.2.5 MIX Method

It is interesting to combine keywords and keyterms and see whether the mixed list of key phrases will bring in more benefit compared to using either keywords or keyterms

alone in key sentence extraction. Our MIX method works as follows:

1. Normalize the weights of 25 keywords to unit length. Do the same for 25 keyterms.

2. Combine 25 keywords and 25 keyterms to obtain a single list of 50 key phrases. In particular, the weight of each keyterm is assigned a factor $\lambda$, i.e., the new weight is $\lambda \cdot W_{Keyterm_i}$.

3. Our objective is to investigate whether keyterms should be given more weight than keywords when they are combined, i.e., determining $\lambda < 1$, $\lambda = 1$, or $\lambda > 1$. We experimented with various values of $\lambda$ and found the best empirical value is 1.5 in terms of the acceptable percentage.

4. Sort the list of 50 key phrases with new weights and select the top 25 key phrases.

5. Re-normalize the new weights of the top 25 key phrases.

## 3.3 Experiments and Evaluation

In this section, we show how summaries of test Web sites are generated, describe the methodology of our user study, and present the evaluation results.

### 3.3.1 Summaries of Test Web Sites

In our work, all five key phrase extraction methods are used to generate key phrases for 20 DMOZ Web sites, which are used in our previous summarization research [85]. These sites are randomly selected from the DMOZ directory because they are either academic or commercial, and users have more familiarity with them. Also these sites are of varying size. The URLs are listed in Table 3.1.

Furthermore, key sentences are extracted from each of the 20 Web sites based on the presence of key phrases [85]. Each Web site summary consists of 25 key phrases and 5 key sentences. A MIX-based summary for the Software Engineering Institute (SEI) Web site[5] is presented in Table 3.2. These summaries are printed out and presented to the human subjects in our user study outlined below.

[5] `http://www.sei.cmu.edu`

Table 3.1: URLs of the 20 test Web sites selected from the DMOZ directory.

| |
|---|
| Software/Software Engineering |
| 1. http://www.ispras.ru/groups/case/case.html |
| 2. http://www.ifpug.org |
| 3. http://www.mapfree.com/sbf |
| 4. http://www.cs.queensu.ca/Software-Engineering |
| 5. http://www.sei.cmu.edu |
| Artificial Intelligence/Academic Departments |
| 6. http://www.cs.ualberta.ca/~ai |
| 7. http://www.ai.mit.edu |
| 8. http://www.aiai.ed.ac.uk |
| 9. http://www.ai.uga.edu |
| 10.http://ai.uwaterloo.ca |
| Major Companies/Publicly Traded |
| 11. http://www.aircanada.ca |
| 12. http://www.cisco.com |
| 13. http://www.microsoft.com |
| 14. http://www.nortelnetworks.com |
| 15. http://www.oracle.com |
| E-Commerce/Technology Vendors |
| 16. http://www.adhesiontech.com |
| 17. http://www.asti-global.com |
| 18. http://www.commerceone.com |
| 19. http://www.getgamma.com |
| 20. http://www.rdmcorp.com |

### 3.3.2 Evaluation Methodology

Evaluation of automatically generated summaries often proceeds in *intrinsic* mode, where summaries are compared against a gold standard, or in *extrinsic* mode, which measures the utility of summaries in performing a particular task (e.g. site browsing).

In this thesis, we aim to investigate how well different types of summaries reveal the main contents of a given Web site[6]. In other words, we are interested in the correctness and completeness of the automatically generated summaries. Our assumption is that the subjects can define the most essential topic of a given Web site well enough for the most essential topic to be used as gold standard. To do so,

[6]We acknowledge that there are other critical factors in multi-document summarization such as coherence, redundancy deduction, and compression rate, which we leave for future research.

Table 3.2: A MIX-based summary for the Software Engineering Institute Web site, consisting of 25 key phrases and 5 key sentences.

| **Part I. top 25 key phrases** |
|---|
| engineering institute, software engineering institute, software engineering, system, software, product line, product, information, software architecture, carnegie mellon university, organization, architecture, capability maturity, institute, program, course, research, carnegie, capability maturity model, defense, development, team, department, term, component |
| **Part II. top 5 key sentences** |
| 1. The Software Engineering Institute (SEI) is a federally funded research and development center sponsored by the U.S. Department of Defense and operated by Carnegie Mellon University.<br>2. The online version of the Annual Report of the Software Engineering Institute (SEI), reporting on fiscal year 2002, is available at `http://www.sei.cmu.edu/annual-report/`.<br>3. The Software Engineering Institute (SEI) sponsors, cosponsors, and is otherwise involved in many events throughout the year.<br>4. The Software Engineering Institute offers a number of courses and training opportunities.<br>5. The Software Engineering Institute (SEI) helps organizations and individuals to improve their software engineering management practices. |

we conducted a user study where summaries are judged by subjects using a golden standard of their own.

### Study Design

We conduct a user study in a "within-subjects" fashion where human subjects read and rate all five summaries of a given Web site (in sheets of paper) based on their understanding of how these summaries relate to the most essential topic of the target Web site. Our study is close to the intrinsic evaluation in the sense that human subjects rate the summaries against a hypothetical gold standard of their own. The study makes sense in that Web site summaries are expected to reveal the main contents of Web sites. Similar studies in which human subjects rate documents or phrases have been reported in [35, 46, 54, 77].

In our study, we focus on the "method" factor only. Other factors such as "subject" (inter-rater reliability) and "Web site" (e.g. academic vs. commercial) might also play a role in this learning task. Inter-rater reliability measures the rating agreement between subjects in a user study. It often calculates a score of how much consensus there is in the ratings given by subjects. There are a number of statistics that can be used to determine the inter-rater reliability. For example, the *joint-probability of agreement* is a simple measure, which takes the number of times each rating (e.g. 1, 2, ..., 5) is given by each subject and then divides this number by the total number of ratings [80]. Investigation of these factors is a topic of future research.

For each given Web site, subjects are asked to do the following:

1. Browse the Web site and extract *the most essential topic*, which is defined as *the entity behind the Web site and its main activity.* The most essential topic serves as the representation of core contents of the target Web site. For example, the most essential topic for the SEI Web site could be extracted as "Software Engineering Institute at CMU for improvement of software engineering management and practice".

2. Read each of the five summaries of the target Web site, which are generated based on the five key phrase extraction methods, respectively.

3. Based on the *relatedness*, which is defined as *the extent to which a summary element (key phrase or key sentence) is related to the most essential topic*, rate summary elements using a 1-to-5 scale (1 = not related, 2 = poorly related, 3 = fairly related, 4 = well related, and 5 = strongly related).

We note that there are several "effects" such as fatigue and practice (warm-up) that could lead to "systematic bias", which means subjects give bias to a particular type of summary. One way to prevent such bias is to randomize the order in which five different summaries of a Web site are presented to subjects. More specifically, for each subject we choose 10 different presentation orders out of 120 possible permutations of five summaries such that the five summaries for each of the 10 Web sites are presented in a different order.

### Study Recruitment

To decide on the number of human subjects, we consulted related studies in the literature. A related research reported in [8] asks 15 subjects to evaluate five summarization methods by collecting data such as number of pen movements in the task of browsing Web pages using handheld devices.

In another study [35], 37 subjects are asked to rate Web pages, which are returned by three different search engines, into "bad", "fair", "good", and "excellent" in terms of their utility in learning about the search topic. However, no specific statistical analysis methods are reported in these two studies.

In [53], 45 subjects are divided into four groups to perform task-based evaluation of multi-document summaries in order to determine whether multi-document summaries measurably improve user performance when using online news browsing systems for directed research.

A size of 20 subjects is sufficient for our study. Each subject is asked to review 10 out of 20 Web sites such that each Web site is covered by exactly 10 subjects. This means that for each method, we have a sample size of 200 with replication.

Participants are graduate students in computer science with strong reading comprehension skills and Web browsing experiences. They are recruited because of the technical nature of the Web sites being summarized. Each subject is provided a computer with Internet access and summaries in hard copies. They are required to finish the study in a session of two hours.

### 3.3.3 Summary Evaluation

In this subsection, we explain how to compare the quality of key phrases and key sentences obtained by different methods based on statistical analysis of rating data collected in the user study. Our main objective is to benchmark the five key phrase extraction methods and investigate which method yields a Web site summary with the best quality.

For each key phrase extraction method, we have a sample size of 200 with replication. Let $n_1$, $n_2$, $n_3$, $n_4$, and $n_5$ be the number of summary elements that receive a score of 1, 2, 3, 4, and 5, respectively. Hence for each summary, $\sum_{i=1}^{5} n_i$ will be 25 for key phrases and 5 for key sentences, respectively.

### Comparison of Key Phrases

We aim to evaluate and compare the five key phrase extraction methods by an analysis of both acceptable percentage and quality value, which are both calculated based on the rating data obtained in the user study.

**Analysis of Acceptable Percentage** Related research in [77] defines acceptable key phrases as those that are rated good or fair by human subjects. In our work, acceptable key phrases and key sentences are those that receive a score of 3, 4, or 5. These summary elements are reasonably related to the most essential topic of a given Web site. In other words, they correctly and completely reveal the main contents of the target Web site. The percentage, $P$, is then formally defined as:

$$P = \frac{n_3 + n_4 + n_5}{\sum_{i=1}^{5} n_i}. \tag{3.6}$$

The five methods TFIDF, KEA, KWD, CNC, and MIX achieve an average acceptable percentage of 0.55, 0.67, 0.59, 0.78, and 0.72, respectively. This indicates that the five methods can be ranked as TFIDF, KWD, KEA, MIX, and CNC, in ascending order of acceptable percentage of key phrases.

To test whether there is a statistically significant difference between the methods, we choose the One-Way Fully Repeated Measures ANOVA, which is a generalized version of $t$-test with replication. Since each method is repeated with multiple Web sites, ANOVA is more appropriate for this analysis.

We apply the One-Way Fully Repeated Measures ANOVA on the acceptable percentage data and a statistically significant difference between the five methods ($F_{4,995} = 23.421$, $P_{value} = 1.58E^{-18}$) is found at the 5% significance level. The $P_{value}$ is approximately zero, which indicates that it is practically certain there is a significant difference between some pair of methods. Hence, we apply ANOVA on each pair of the five methods. The ANOVA results are presented in Table 3.3, which can be summarized as TFIDF < KWD << KEA < MIX << CNC and KEA << CNC[7].

**Analysis of Quality Value** In addition to the acceptable percentage measure, we also aim to compare the five methods using the *quality value* measure, which calculates

[7] << indicates a statistically significant difference with $P_{value} \leq 0.05$; < indicates a difference with $P_{value} > 0.05$. Moreover, << or < is not transitive in the mathematical sense.

Table 3.3: $F_{1,398}$ and $P$ values of applying ANOVA on each pair of the five key phrase extraction methods, i.e., TFIDF, KEA, KWD, CNC, and MIX, using the measure of acceptable percentage of key phrases.

| Method | KEA | KWD | CNC | MIX |
|---|---|---|---|---|
| TFIDF | $F = 19.121$<br>$P = 1.57E^{-5}$ | $F = 2.224$<br>$P = 0.137$ | $F = 72.495$<br>$P = 3.47E^{-16}$ | $F = 37.071$<br>$P = 2.68E^{-9}$ |
| KEA | | $F = 8.048$<br>$P = 0.005$ | $F = 17.765$<br>$P = 3.09E^{-5}$ | $F = 3.051$<br>$P = 0.082$ |
| KWD | | | $F = 48.312$<br>$P = 1.50E^{-11}$ | $F = 20.634$<br>$P = 7.38E^{-6}$ |
| CNC | | | | $F = 6.079$<br>$P = 0.014$ |

the average correctness score of summary elements. The quality value, $Q$, of 25 key phrases in a summary is defined as follows:

$$Q = \frac{\sum_{i=1}^{5} n_i \times i}{\sum_{i=1}^{5} n_i}. \tag{3.7}$$

The higher the quality value, the more accurately the summary reveals the main contents of a site overall.

The acceptable percentage measure and the quality value measure are intrinsically related to each other as they are both based on users' ratings. The only difference is that the former gives equal weight to (i.e., a summation of) the number of summary elements with scores 3, 4, and 5, while the latter gives different weights to summary elements with different scores (i.e., number of such elements times the score they receive).

The average quality values of key phrases extracted by TFIDF, KEA, KWD, CNC, and MIX, are 2.85, 3.55, 3.46, 3.96, and 3.87 out of a possible 5.0, respectively. Hence the ordering of methods in terms of quality values is exactly the same as that obtained by the acceptable percentage measure. We also apply ANOVA on the quality value data. We obtain the same result as using the acceptable percentage measure with the only exception that there is no statistically significant difference between MIX and CNC, i.e., MIX < CNC.

### Comparison of Key Sentences

In our Web site summarization framework [85], once key phrases are identified by a particular method, we further extract key sentences based on the density of key phrases. We are interested in learning how good key sentences, which are obtained by using different key phrase extraction methods, will be from the user's point of view. Again, we are using both the acceptable percentage and quality value measures introduced in Equations 3.6 and 3.7, respectively.

The key sentences resulted from the five methods TFIDF, KEA, KWD, CNC, and MIX achieve an average acceptable percentage of 0.88, 0.90, 0.89, 0.90, and 0.91, respectively. The One-Way Fully Repeated Measures ANOVA on the acceptable percentage data shows that there is no statistically significant difference between the five methods ($F_{4,995} = 0.490$, $P_{value} = 0.743$).

However, we note that compared with the ordering of key phrase extraction, KEA and MIX have moved up in the ordering of key sentence extraction, i.e., KEA is tied with CNC compared to that KEA is worse than CNC in key phrase extraction, and MIX is better than CNC compared to that MIX is worse than CNC in key phrase extraction. This indicates that a mixture of single keywords and multi-word keyterms can improve the key sentence extraction performance. This also implies that key sentence extraction is often dominated by a few "good" key phrases, which are often ranked high in the key phrase list. Moreover, we observe that on average any two methods share 2.3 out of 5 key sentences.

The average quality values of key sentences resulted from TFIDF, KEA, KWD, CNC, and MIX, are 3.87, 3.99, 3.94, 4.01, and 4.02, respectively. Hence the ordering of methods in terms of quality values is similar with that obtained by the acceptable percentage measure, i.e., the MIX method is the best in terms of key sentence extraction. The ANOVA test shows that there is no statistically significant difference between methods ($F_{4,995} = 1.145$, $P_{value} = 0.334$).

### Comparison of Summaries

Each summary consists of 25 key phrases and 5 key sentences, so the evaluation of the whole summary depends on users' relative preference to different parts of the summary. A simple survey in our study indicates that users prefer to give equal weight

to both parts of the summary. Thus the quality value of a summary will be $\frac{Q_p+Q_s}{2}$, where $Q_p$ and $Q_s$ are quality values of key phrases and key sentences (calculated using Equation 3.7), respectively. The ANOVA based on summary quality values shows that TFIDF < KWD << KEA < MIX < CNC and KEA << CNC.

A thorough user study is needed in future research to see what is the best size for summary elements and how the methods compare with each other when the size of summaries changes.

#### Comparison of Computational Cost

Regarding the computational complexity, it is observed that on average, TFIDF, KWD and KEA are roughly 12 times faster than CNC in extracting key phrases from the narrative text of a given Web site. CNC is much slower mainly due to the computational complexity of NC-value. Hence in terms of summary quality, CNC is the best choice and MIX is a good alternative in the automatic Web site summarization task, whereas if efficiency is the most important factor, then KEA is the best method.

### 3.3.4 Summary

We evaluate the correctness and completeness of summary elements in an intrinsic manner, i.e., we measure how well different types of summaries could reveal the core contents of Web sites. We are interested in learning why and in what circumstances one method outperforms the other in this task.

It is not surprising that TFIDF is the worst method as it is conceptually simple to consider only features of term frequency and document frequency. The KWD method is able to take advantage of topical information in three categories of text. However, it is mainly based on analysis of a word's overall frequency in the document collection. Consequently, its performance is at the same level as TFIDF. The KEA method utilizes both the TFIDF feature and the first appearance feature. It provides a learning scheme where prior knowledge of key phrases can be easily incorporated as the learning model is conceptually domain-independent. Hence, it can find a better set of key phrases than TFIDF and KWD. The CNC method incorporates both statistical information (frequency, term nesting statistic, and contextual information)

and linguistic knowledge. Consequently, it is able to find the best set of key phrases. Finally, the MIX method has the advantage of obtaining a good mixture of single keywords and multi-word keyterms, which are found to greatly improve the performance of key sentence extraction.

KWD and KEA are supervised methods which require known phrases from training documents in order to obtain the model. In contrast, TFIDF and CNC are unsupervised methods where no learning process is involved. Hence, they are more practical when applied to applications without domain knowledge. However, the CNC method is more sensitive to the amount of narrative text than the other three methods as it prefers more narrative text to conduct the *NC-value* calculation.

It will be ideal to apply the MIX method to obtain a good set of candidate key phrases which can be further processed in consideration of Web-specific features such as availability of phrases in meta data and anchor text. Also more advanced learning algorithm such as Support Vector Machines can be deployed. This will be a direction of our future research.

# Chapter 4

# Web Page Clustering

In the previous chapter, we investigate five key phrase extraction methods in the straightforward summarization system and demonstrate that CNC is the best method. In this chapter, we investigate the clustering problem in our clustering-summarization framework. We use both text- and link-based features to find in a given Web site the most essential topics, which will be further summarized using the CNC-based summarization system to create multiple short cluster summaries.

## 4.1 Introduction

Text clustering has been extensively researched in many text-based applications. It plays an important role in organizing large document collections into a small number of meaningful topic groups [65]. A variety of approaches (e.g. [4, 15, 30, 72]) to text clustering have been developed. Typically clustering approaches can be categorized as *agglomerative* or *partitional* based on the underlying methodology of the algorithm, or as *hierarchical* or *flat* (non-hierarchical) based on the structure of the final solution [87].

In general, text clustering involves constructing a vector space model and representing documents by feature vectors. First, a set of features is properly selected from the document corpus. Second, each document is represented by a feature vector, which consists of the weights of all chosen features. Finally, clustering proceeds by measuring the similarity (e.g. a function of Euclidean distance) between documents and assigning documents to appropriate clusters.

Web page clustering, which has recently been of significant interest in the Web content mining field [43], is a useful tool for summarization, organization and navigation of semi-structured Web documents [79]. For instance, an effective clustering system can help greatly improve the organization and presentation of search engine results.

Approaches to Web page clustering have been either *text-based* or *link-based*. Text-based approaches [7, 12] use a set of common terms shared among documents as features. Due to the large number of words in the vocabulary, this approach typically results in a very high dimensional document representation. On the other hand, link-based approaches [13, 78] analyze the hyperlinks between Web documents for feature selection. The feasibility of such approaches depends on availability of a rich set of hyperlinks. Some Web page clustering systems [56, 79] use a combination of the above two.

In this chapter, we investigate the problem of Web page clustering in the context of Web site summarization. In order to effectively summarize an entire multi-topic Web site, we need to find the essential topics in a given Web site. Thus, we perform a coupled text- and link-based clustering on Web pages from the target Web site to find meaningful topic groups, which can further be individually summarized by a summarization system [85].

However, in text-based clustering where the bag-of-words representation is used, a very high dimensional feature space is often required. These features may not be equally useful. Noise words may not contribute to or even degrade the clustering process. Thus, the task of selecting the "best" feature subset, known as Feature Selection (FS), is an important task[1].

Feature selection has been well studied in text categorization [82] and text clustering [45] tasks (a full review of the literature can be found in [44]). Since Web pages often contain more noise (e.g. navigational menu) than traditional plain text, it is more important to perform proper feature selection on the text part. In this thesis, we investigate whether standard feature selection methods, including document frequency, term variance, information gain, mutual information, and $\chi^2$ statistic, can improve the quality of text-based Web page clustering. For the link-based clustering, we apply co-citation and bibliographic coupling to learn whether linkage information can improve the clustering quality.

More specifically, we apply $K$-means and $X$-means [58] (an extension of $K$-means that identifies the optimal number of clusters within a given range) algorithms to

[1]Note that feature selection is different from feature extraction in the sense that the former chooses an optimal subset of features while the latter often introduces a new and smaller feature space via projection or mapping.

perform clustering on a set of Web pages from a given Web site. The clustering quality is evaluated using entropy and accuracy.

## 4.2 Web Page Corpora

In order to perform Web page clustering experiments we choose two test Web sites, which are the Software Engineering Institute (SEI) Web site[2], and the Air Canada (AC) Web site[3]. The two Web sites have been extensively used in our previous Web-based research [85, 83, 86]. They are well designed in that most of the Web pages are static HTML files. More importantly, each Web page can be easily labelled into one of a set of topics, which are defined by the Web site designers. The topic information will be used for clustering evaluation purposes (see Section 4.4). In the following subsection, we discuss how to crawl a given Web site and collect various features of Web pages for the clustering purpose.

### 4.2.1 Web Site Traversal

Intuitively, Web site designers often construct a Web site and organize its contents in a hierarchical manner. In a given Web site, each Web page is uniquely identified by its Uniform Resource Locator (URL). The home page often introduces the entity behind the target Web site in a general mode. All Web pages, which are pointed to by the home page, often cover main topics of the Web site. If we go deeper into the Web site, then pages tend to discuss specific topics with more details.

#### Breadth-first Search

In order to effectively summarize a given Web site, we need to crawl the site and obtain its link structure. The breadth-first search algorithm is often used for this purpose. The algorithm has two auxiliary lists. One is $Q$, a queue of pages to be visited from the front end one by one. Initially, $Q$ has only one node, which is the home page. The other is $V$, a list of pages that are already visited. At each level of the breadth-first traversal, each Web page will be tested whether it has been visited and various features will be updated accordingly. The search process continues until

[2] `http://www.sei.cmu.edu`, last crawled on November 15, 2005.
[3] `http://www.aircanada.ca`, English version, last crawled on November 15, 2005.

no more new Web pages can be found, or a desired number of pages or depth levels has been visited. The pseudo code of the algorithm is shown in Figure 4.1.

---

*Input*: URL of the home page of a given Web site.
*Output*: A link hierarchy of the target Web site.
*Initialization*: Mark all nodes unvisited.
**begin**
    put home page into $Q$
    **while** $Q$ is not empty
        $u$ = front element of $Q$
        remove $u$ from Q and add $u$ to $V$
        visit $u$
    **end while**
**end**

---

Figure 4.1: Algorithm of breadth-first Web site traversal.

When a Web page is visited in the Web site traversal, various features are dynamically updated. In our work, we look at both text- and link-based features. The former contains the plain text extracted from the HTML source of a Web page, while the latter mainly consists of incoming and outgoing links of the current page. The features that we aim to investigate for each individual Web page are summarized in Table 4.1.

Table 4.1: Feature list of a Web page.

| Notation | Feature | Meaning |
|---|---|---|
| $u$ | URL | Uniform Resource Locator of a Web page |
| $d$ | depth | depth level in the breadth-first traversal |
| $t$ | plain text | plain text extracted from the HTML source |
| $I$ | incoming links | set of incoming links |
| $O$ | outgoing links | set of outgoing links |

The feature update process consists of items such as calculating the depth level of an unvisited page, updating the incoming and outgoing link sets, etc. The pseudo code, which updates the feature list when a Web page is visited, is presented in Figure 4.2.

*Input*: $Q$, $V$, and Web page $u$.
*Output*: Feature list of various Web pages.
**begin**
  parse Web page $u$ to obtain $t$ and $O$
  **for** each $v \in O$ **do**
    **if** $v \notin V$, **then**
      add $v$ to both $Q$ and $V$ and $d_v = d_u + 1$
    **endif**
    add $v$ to $O(u)$ and add $u$ to $I(v)$
  **end for**
**end**

Figure 4.2: Algorithm of Web page feature update.

For small and medium size Web sites[4], it is feasible to crawl the whole host for complete link hierarchy construction. However, for large Web sites (e.g. `http://www.microsoft.com`), it is very time-intensive to do so. In such cases, the number of levels to crawl should be properly determined in order to construct a link hierarchy which can effectively represent the whole Web site. In our work, The site traversal stops when either a total of 1000 Web pages have been crawled, or the crawler has finished crawling the fourth depth level, whichever comes first. The maximum values 1000 and 4 are empirical numbers set in [85].

### Link Hierarchy

The resulting link structure from the breadth-first traversal forms a directed Web graph, as shown in Figure 4.3. Each node in the link hierarchy represents a Web page, which is uniquely identified by its URL. An arrow indicates that there is a hyperlink pointing from one page to another page. Pages at top levels tend to describe general topics, while pages at bottom levels more often discuss details of these topics.

In such a link hierarchy, there are mainly three types of links.

- **Forward link**: a hyperlink in one Web page pointing to another Web page at a lower layer, as labelled $< 1 >$ in Figure 4.3.

[4]By Web site we mean all Web pages that reside in a unique host. Investigation of out-of-host pages is not our focus in this thesis.

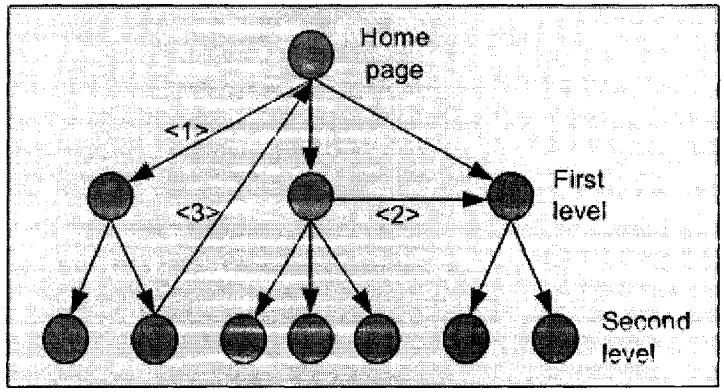


Figure 4.3: Hierarchical link structure of a Web site.

- **Cross link**: a hyperlink in one Web page pointing to another Web page at the same layer, as labelled $< 2 >$ in Figure 4.3.

- **Back link**: a hyperlink in one Web page pointing to another Web page at a higher layer, as labelled $< 3 >$ in Figure 4.3.

These three types of links form the link hierarchy of a given Web site. It is observed that forward-links are of the majority. We aim to investigate the significance of these links in the Web site summarization framework.

In our work, only pages within the target host are collected. Investigation of out-of-host pages is a future research direction. Also only Web pages of the **text/html** content type are kept and all other application files (e.g. .jpg, .gif, and .doc) are removed for simplifying text processing. More details are available in Appendix A.

The above crawling process leads to a smaller set of 927 Web pages for the SEI site and 627 for the AC site. Each Web page left in the collection is manually visited and to which a topic label is assigned. The Web site designers have provided a topic for each Web page. We go through the designer's topic assignment page by page. If we do not agree with the Web site designers, we correct it based on the text and additional information such as the "Keywords" and "Metadata" fields embedded in the HTML code. We observe that more than 95% of the times we agree with the Web site designers about the topic assignment. Only the primary topic is assigned to a Web page when there are multiple topics available. Finally, plain text from all Web pages is extracted by the text browser Lynx, which has been found to outperform several alternative text extraction tools [85] in terms of more effective selection of

plain text.

### 4.2.2 SEI Corpus

The final SEI corpus after preprocessing is summarized in Table 4.2, where $i$ is the topic index and $|T_i|$ is the number of Web pages in each topic $T_i$. The topics are presented in the order of first appearance in the site traversal.

Table 4.2: Topic distribution for the Software Engineering Institute Web site.

| $i$ | Topic $T_i$ | $\|T_i\|$ |
|---|---|---|
| 1 | About SEI | 180 |
| 2 | Software Management | 115 |
| 3 | Software Engineering | 141 |
| 4 | Software Acquisition | 32 |
| 5 | Work With Us | 90 |
| 6 | Software Products | 225 |
| 7 | Publications | 92 |
| 8 | News | 52 |

The most populated topic is *Software Products* with 225 documents while the least populated topic is *Software Acquisition* with only 32 documents, as shown in Table 4.2.

### 4.2.3 AC Corpus

The final AC corpus after preprocessing is summarized in Table 4.3, where $i$ is the topic index and $|T_i|$ is the number of Web pages in each topic $T_i$. Again, the topics are presented in the order of first appearance in the site traversal.

The most populated topic is *News and Media* with 174 documents while the least populated topics are *Investor* and *Aeroplan* with only 21 documents, as shown in Table 4.3.

We observe that both corpora have an imbalanced topic distribution, i.e., a large number of documents are from few topic groups. This may affect feature selection in text-based clustering, i.e., the topic distribution imbalance may favor non-topical methods such as Term Variance while worsening the performance of topical methods such as Information Gain and Mutual Information.

Table 4.3: Topic distribution of the Air Canada Web site.

| $i$ | Topic $T_i$ | $\|T_i\|$ |
|---|---|---|
| 1 | About Air Canada | 96 |
| 2 | News and Media | 174 |
| 3 | Flights | 135 |
| 4 | Hotels | 36 |
| 5 | Cars | 30 |
| 6 | Vacation | 75 |
| 7 | Investor | 21 |
| 8 | Career | 39 |
| 9 | Aeroplan | 21 |

## 4.3 Web Page Clustering

In this section we first describe the clustering algorithms used in this work: $K$-means and $X$-means. Then we explain how to represent Web pages with text- and link-based features. We also give a brief description of five text-based feature selection methods with which we choose to experiment.

### 4.3.1 Clustering Algorithms

In this thesis, we experiment with both $K$-means and $X$-means algorithms. $X$-means is an extension of the conventional $K$-means algorithm. Our goal is to investigate whether basic clustering algorithms can effectively detect topic groups for further summarization. Investigation of advanced clustering techniques is one of the future research directions.

#### $K$-means

The conventional $K$-means algorithm has been widely used in clustering tasks due to its implementation simplicity and low computational complexity. This non-hierarchical method first selects $K$ data points using a random seed. The data points are used as the initial centers for the $K$ clusters, one for each cluster. Second, all data points are assigned to the cluster whose centroid is the closest (e.g. in terms of Euclidean distance). Third, the centroid of each cluster is recalculated based on the points

assigned to it. Steps two and three are repeated until the centers do not change [72].

### $X$-means

One major shortcoming of $K$-means is that the number of clusters, $K$, has to be provided beforehand, which is often difficult to decide without any prior topical knowledge of a given document corpus. $X$-means [58], an extension to the standard $K$-means, has the advantage of estimating the optimal number of clusters within a given range. This is useful in finding topic groups in a given Web site.

$X$-means is a variation of the $K$-means algorithm. After each run of $K$-means, it computes the Bayesian Information Criterion [58] to decide which subset of the current centers should be further split in order to better fit the data. The experiments reported in [58] show that $X$-means produces better clustering quality than a more traditional method that estimates the number of clusters by guessing $K$.

In this work, we experiment with both $K$-means and $X$-means to perform Web page clustering. The implementations of $K$-means and $X$-means are from the WEKA[5] software package and the $X$-means' authors[6], respectively[7]. Both implementations use the Euclidean distance to measure the similarity between two documents, i.e., the smaller the distance is, the more similar they are.

## 4.3.2 Document Representation

One important task in the clustering process is how to represent documents. The Vector Space Model is often used for this purpose. Each document is represented by a vector which consists of a set of features that are properly chosen from the feature space.

In this work, we look at both text- and link-based features. Text-based features include a list of key phrases that are extracted from the text body of Web pages using automatic phrase extraction methods. In this work, we look at three methods, TFIDF, KEA and CNC. Link-based features consist of incoming links and outgoing links, which are inspired by the ideas of co-citation [68] and bibliographic coupling [34] in the scientific literature, respectively.

[5] `http://www.cs.waikato.ac.nz/ml/weka`

[6] `http://www.cs.cmu.edu/~dpelleg/kmeans.html`

[7] We are very thankful to the authors for providing us the code.

### Text-based Features

For each Web page in the above two Web corpora, we obtain its plain text using Lynx. Then we separately apply three key phrase extraction methods, TFIDF, KEA, and CNC to obtain phrases for document representation.

**TFIDF** TFIDF has been widely used as bag-of-words representation in clustering applications. In this work, we use TFIDF as a baseline method. Keyword identification involves in the following steps:

First, a standard list of 425 stopwords [20] is removed from each text file (plain text of a Web page). Second, plain text is tokenized and each unique word is stemmed using the Porter stemmer, and its frequency in the current text file (TF part) and the number of documents where it appears (DF part) are recorded. After all the documents are scanned, $tfidf$ values of words (more precisely, word stems) in each document are calculated and normalized as in Equations (3.1) and (3.2), respectively.

All the unique terms in the document collection are ranked according to their $df$ values. Those terms with a $df$ value above a certain threshold form the feature space for document representation.

**KEA** Automatic Keyphrase Extraction (KEA) [81] is an efficient and practical algorithm for extracting key phrases from a document corpus. It consists of two stages: training and extraction.

We apply the NTXT model built in Subsection 3.2.2 to the Web pages. All the candidates with a probability above a certain threshold form the feature space for document representation. Their $tfidf$ values in each document are calculated and normalized as in Equations (3.1) and (3.2), respectively.

**CNC** C-value/NC-value (CNC) is a domain-independent method used to automatically extract multi-word phrases from the whole document corpus.

C-value is a measure of term likelihood calculated for each candidate term based on its frequency in the corpus and the frequency of its occurrence as nested within longer terms.

NC-value is an extension to the C-value, incorporating information of context

words into term likelihood. Context words are those that appear in the vicinity of candidate phrases, i.e., nouns, verbs and adjectives that either precede or follow the candidate phrase.

The final phrase list is ranked by *NC-value*. All phrases above a certain pre-defined threshold form the feature space for document representation. Their $tfidf$ values in each document are calculated and normalized as in Equations (3.1) and (3.2), respectively.

### Link-based Features

For the link analysis, incoming and outgoing links of each Web page are recorded during the Web site traversal.

In the scientific literature, co-citation [68] refers to the case that two documents $d_1$ and $d_2$ are cited by a common third document $d_3$. The more co-citations the two documents share, the higher the co-citation strength is. On the other hand, bibliographic coupling [34] occurs when two documents $d_1$ and $d_2$ cite a common third document $d_3$. Similarly, the more citations the two documents have in common, the higher the coupling strength is. Hence, these two measures are widely used to estimate the similarity between two documents in a hypertext context.

In the Web context, co-citation occurs when two Web pages $p_1$ and $p_2$ are pointed to by a common third page $p_3$, while bibliographic coupling happens when two Web pages $p_1$ and $p_2$ both point to a common third page $p_3$.

Recall that in the Web site crawling stage, the breadth-first search algorithm is used and only pages within the target host are crawled and collected. Each unique Web page is assigned a *depth* value when it appears in the site traversal for the first time and the home page has a depth of 0. All the crawled pages are sorted in the order of *first appearance* in the site traversal. Those pages whose depth value are lower than a pre-defined threshold form the link-based feature space. The link-based vector for each document consists of binary numbers. If a page in the feature space appears as an incoming or outgoing link of the current document, then the corresponding entry in the link-based vector is 1, otherwise 0.

The above link analysis has been a prominent tool in many fields such as Web

information retrieval and Web mining. However, hyperlinks in the Web are not as organic as references between research articles in the scientific literature. Consequently, we may have to rely more heavily on text-based features and investigate whether incorporation of link-based features can improve the clustering quality.

### 4.3.3 Feature Selection Methods

Clustering of documents often suffers from high dimensionality of the feature space if the bag of words representation is used. For example, when the TFIDF method is used, there are as many as tens of thousands of unique words in the corpus and many of them are noise which has no discrimination power against documents. Feature selection involves ranking the feature list and choosing a particular subset of features to represent documents. The subset could be chosen in various ways, for instance, the top $k$ features, or features with a score of more than a pre-determined threshold.

In this work, we investigate five text-based feature selection methods. They are Document Frequency (DF), Term Variance (TV), Information Gain (IG), Mutual Information (MI), and the $\chi^2$ statistic (CHI). The first two methods do not need any information about the topic distribution of the document collection, so they are unsupervised methods. In the contrast, the last three methods require external topical knowledge, so they are supervised methods.

In all methods, let $N$ be the total number of documents in the corpus.

#### Document Frequency

Document frequency is the number of documents in which a term (more precisely, its stem) appears. It is a simple and popular metric to measure a term's popularity of presence in the global corpus. Let $n$ be the document frequency, of a term $t$. Hence, $n \in [1, N]$.

For each unique term in the corpus, we compute its document frequency and remove all terms whose document frequency is less than a pre-defined threshold. The underlying assumption is that terms that are too rare are not informative and thus could be removed to reduce the feature space [82].

### Term Variance

Similar with document frequency, term variance [36] is another simple topic-free metric. It measures the variance of a term's frequency of occurrence in all documents. The variance $V$ of a term $t$ is formally defined as follows:

$$V(t) = \sum_{i=1}^{N} f_i^2(t) - \frac{1}{N}(\sum_{i=1}^{N} f_i(t))^2. \tag{4.1}$$

where $f_i(t)$ is the number of times that term $t$ appears in document $i$.

For each unique term in the corpus, we compute its term variance and remove all terms whose term variance is less than a pre-defined threshold. The underlying reasoning is that if a term's occurrence is evenly distributed over all documents, then it has little power to discriminate documents. Hence, the quality of a term is proportional to its term variance score, i.e., the higher the $V(t)$ score, the better the term is.

### Information Gain

Information gain [82] is a term goodness criterion commonly used in the text categorization task. It measures the number of bits of information obtained for topic prediction given the knowledge of presence and absence of a term in a document. The information gain $G$ of a term $t$ is formally defined as follows:

$$\begin{aligned} G(t) &= -\sum_{i=1}^{l} P(T_i) \log P(T_i) \\ &\quad + P(t) \sum_{i=1}^{l} P(T_i|t) \log P(T_i|t) \\ &\quad + P(\bar{t}) \sum_{i=1}^{l} P(T_i|\bar{t}) \log P(T_i|\bar{t}). \end{aligned} \tag{4.2}$$

where $l$ is the number of topics in the given corpus; $P(T_i)$ is the fraction of documents with topic $T_i$, i.e., $|T_i|/N$; $P(t)$ is the fraction of documents where term $t$ appears, i.e., $n/N$; $P(T_i|t) = m/n$ and $m$ is the number of documents with topic $T_i$ where term $t$ appears; $P(\bar{t}) = 1 - n/N$; and $P(T_i|\bar{t}) = \frac{|T_i|-m}{N-n}$.

For each unique term in the corpus, we compute its information gain and remove all terms whose information gain is less than a pre-defined threshold. The underlying

reasoning is that terms with high information gain are useful for topic prediction. Hence, the quality of a term is proportional to its information gain score, i.e., the higher the $G(t)$ score, the better the term is.

### Mutual Information

Mutual information [82] is a term goodness function often used in statistical language modelling of word associations with topics. Intuitively, it measures the information that a term and the topics share. For example, if term $t$ and topic $T_1$ are independent, then knowing $t$ does not give any information about $T_1$ and vice versa. Hence, their mutual information is zero. At the other extreme, if term $t$ and topic $T_1$ are identical then all information conveyed by $t$ is shared with $T_1$, i.e., knowing $t$ determines $T_1$ and vice versa. The mutual information $I$ of a term $t$ is formally defined as follows:

$$I(t) = \sum_{i=1}^{l} P(T_i|t) \log \frac{P(t|T_i)}{P(t)}. \tag{4.3}$$

where $P(T_i|t) = m/n$; $n$ is the document frequency of term $t$; $m$ is the number of documents with topic $T_i$ where term $t$ appears; and $P(t|T_i) = m/|T_i|$.

For each unique term in the corpus, we compute its mutual information with all topics and remove all terms whose mutual information is less than a pre-defined threshold. The underlying reasoning is that terms with high mutual information have more dependence with topics. Hence, the quality of a term is proportional to its mutual information score, i.e., the higher the $I(t)$ score, the better the term is.

### $\chi^2$ Statistic

The $\chi^2$ statistic [82] can be used as a term goodness function to measure the lack of independence between a term and a topic and can be compared to the $\chi^2$ distribution with one degree of freedom. The $\chi^2$ of a term $t$ can be formally defined as follows:

$$\chi^2(t) = \sum_{i=1}^{l} P(T_i) \frac{N \cdot (P(t,T_i) \cdot P(\bar{t},\overline{T_i}) - P(t,\overline{T_i}) \cdot P(\bar{t},T_i))^2}{P(t) \cdot P(\bar{t}) \cdot P(T_i) \cdot P(\overline{T_i})}. \tag{4.4}$$

where: $P(t,T_i) = m/N$ and $m$ is the number of documents with topic $T_i$ where term $t$ appears; $P(\bar{t},\overline{T_i}) = 1 - \frac{n+|T_i|-m}{N}$ and $n$ is the document frequency of term $t$; $P(t,\overline{T_i}) = \frac{n-m}{N}$; and $P(\bar{t},T_i) = \frac{|T_i|-m}{N}$.

For each unique term in the corpus, we compute its $\chi^2$ statistic and remove all terms whose $\chi^2$ statistic is less than a pre-defined threshold. The underlying reasoning is that terms with high $\chi^2$ statistic have more dependence with topics. Hence, the quality of a term is proportional to its $\chi^2$ statistic score, i.e., the higher the $\chi^2(t)$ score, the better the term is.

Document frequency and term variance have a linear computational complexity in terms of number of terms in all documents, while information gain, mutual information, and $\chi^2$ statistic have a quadratic computational complexity. Moreover, if $m = 0$, i.e., term $t$ and topic $T_i$ are independent, then the corresponding part in Equations 4.2, 4.3, and 4.4, has a natural value of 0.

## 4.4 Experiments and Evaluation

In this section we discuss our experiments of text-based, link-based, and coupled text- and link-based clustering. In each part, we show results of clustering experiments and statistical tests. Our main objective is to learn the influence of feature selection on text-based clustering. Additionally, we are interested in learning whether link-based features can improve the clustering quality.

### 4.4.1 Evaluation Schemes

Evaluation of a particular clustering often uses either *internal quality measure* or *external quality measure*. Internal quality measure maximizes the overall similarity within clusters and dissimilarity between clusters without reference to external topical knowledge. On the other hand, external quality measure such as *entropy* and *F*-measure examines the clustering quality by comparing the resulting clusters to known topic memberships [72].

In this work, we use *entropy* and *accuracy* to evaluate the quality of a particular clustering result $C = \{C_1, C_2, ..., C_k\}$ with respect to known topics $T = \{T_1, T_2, ..., T_l\}$. Each cluster $C_i$ $(1 \leq i \leq k)$ or topic $T_j$ $(1 \leq j \leq l)$ represents a set of documents.

### Entropy

Entropy [55] measures the purity or uniformity of clusters with respect to known topics. It is formally defined as the weighted sum of entropies for all clusters as shown in Equation 4.5. The smaller the entropy, the better the result.

$$E(C) = -\sum_{i=1}^{k} \frac{|C_i|}{N} \cdot \sum_{j=1}^{l} p_{i,j} \log p_{i,j}. \tag{4.5}$$

where $p_{i,j}$ is the probability that a document in cluster $C_i$ is of the topic $j$, estimated by $\frac{|C_i \cap T_j|}{|C_i|}$.

### Accuracy

Accuracy (also known as precision) [45] is an intuitive method to calculate the average quality of all clusters. It is formally defined as the weighted sum of accuracies for all clusters as shown in Equation 4.6.

$$A(C) = \sum_{i=1}^{k} \frac{|C_i|}{N} \cdot \frac{\max_{j=1}^{l} |C_i \bigcap T_j|}{|C_i|}. \tag{4.6}$$

which is equivalent to

$$\frac{1}{N} \sum_{i=1}^{k} \max_{j=1}^{l} |C_i \bigcap T_j|. \tag{4.7}$$

### 4.4.2 Text-based Clustering

For text-based clustering, we are interested in finding out whether there is a statistically significant difference between the quality of the following clustering options:

- Clustering methods: $K$-means vs. $X$-means. In $K$-means clustering, $K$ represents the desired number of clusters and has to be pre-defined. In our case, the topic groups are already known. Hence, we set $K$ equal to the known number of topics in each Web corpus, i.e., 8 for the SEI corpus and 9 for the AC corpus. For $X$-means clustering, we let $X$-means determine the optimal number of clusters within the range of [1, 20].

- Document representation: TFIDF vs. KEA vs. CNC. We separately apply TFIDF, KEA, and CNC to obtain a bag of phrases and define feature sets

using the feature selection methods. For example, when the KEA method is used, the sets of phrases from each web page are united and then phrases are ranked by each feature selection method. Thus we have five different feature sets for each type of text-based document representation.

- Feature selection: DF vs. TV vs. IG vs. MI vs. CHI. For each document representation used, the five feature selection methods are separately used to define feature sets.

- Dimensionality reduction: we perform clustering in both high and low dimensional space to see if feature selection methods can reduce the dimensionality while maintaining the quality of clusters. The eight different dimensionalities we choose are: 50, 100, 200, 300, 500, 1000, 2000, and 3000. For example, when the dimensionality of 1000 is used, the top 1000 phrases of each feature list will be used to represent all documents using the normalized $tfidf$ values of selected phrases.

We enumerate all configurations of the above options to evaluate text-based clustering. This leads to a total of 2 (algorithms) $\times$ 3 (document representations) $\times$ 5 (feature selections) $\times$ 8 (dimensionalities) = 240 clustering configurations for each Web corpus. We denote each clustering configuration in the order of *clustering method*, *document representation*, *feature selection*, and finally *dimensionality*. For instance, the configuration of *$X$-means clustering with 500 KEA phrases ranked by TV* will be denoted as *xm-kea-tv-500*.

For each configuration, the clustering is repeated for 20 times using 20 randomly chosen seeds. The 20 repeated runs produce a list of 20 entropy and 20 accuracy values for each clustering. The mean entropy (denoted as $\overline{e}$), or the mean accuracy (denoted as $\overline{a}$) over all 20 runs is taken as the **quality** of this particular clustering. As an example, Table 4.4 shows the entropy and accuracy values for configurations *km-kea-tv-500* and *xm-kea-tv-500* on the AC corpus.

As we can see in Table 4.4, $K$-means and $X$-means achieve a mean entropy of 0.9396 and 0.7312, respectively, and a mean accuracy of 0.6388, and 0.6983, respectively. This indicates that $X$-means algorithm produces a clustering with higher quality than $K$-means does.

Table 4.4: Entropy and Accuracy values for $K$-means and $X$-means clustering on the AC corpus using 500 KEA phrases ranked by TV.

| | *Entropy* | | *Accuracy* | |
|---|---|---|---|---|
| Run | $K$-Means | $X$-Means | $K$-Means | $X$-Means |
| 1 | 1.1112 | 0.7468 | 0.6029 | 0.7081 |
| 2 | 0.8396 | 0.8578 | 0.6746 | 0.6699 |
| 3 | 1.0502 | 0.7873 | 0.5981 | 0.6555 |
| 4 | 0.8910 | 0.9596 | 0.6364 | 0.6411 |
| 5 | 0.9587 | 0.7730 | 0.6411 | 0.6746 |
| 6 | 0.9469 | 0.5581 | 0.6459 | 0.7751 |
| 7 | 0.7716 | 0.6094 | 0.7129 | 0.7321 |
| 8 | 1.0240 | 0.5844 | 0.6029 | 0.7656 |
| 9 | 0.7863 | 0.6070 | 0.6842 | 0.7512 |
| 10 | 0.9247 | 1.0141 | 0.6077 | 0.5885 |
| 11 | 0.9947 | 0.7589 | 0.5837 | 0.6890 |
| 12 | 1.0732 | 0.7789 | 0.6316 | 0.7033 |
| 13 | 1.0223 | 0.8131 | 0.6220 | 0.6555 |
| 14 | 0.8415 | 0.6545 | 0.6603 | 0.7129 |
| 15 | 0.7768 | 0.6314 | 0.6938 | 0.7512 |
| 16 | 0.8671 | 0.6419 | 0.6651 | 0.7368 |
| 17 | 0.8886 | 0.6452 | 0.6699 | 0.7033 |
| 18 | 0.8298 | 0.6425 | 0.6555 | 0.7177 |
| 19 | 0.9630 | 0.7116 | 0.6172 | 0.7033 |
| 20 | 1.2316 | 0.8484 | 0.5694 | 0.6316 |
| $\overline{e}/\overline{a}$ | 0.9396 | 0.7312 | 0.6388 | 0.6983 |
| $s_d$ | 0.1221 | 0.1252 | 0.0385 | 0.0483 |

### Comparison of All Configurations

We are interested in finding out which clustering configuration can lead to the best clustering quality. For each Web corpus, we sort all the configurations in ascending order of mean entropy over 20 runs (the lower, the better) and in descending order of mean accuracy over 20 runs (the higher, the better), respectively.

**Results on the SEI Corpus** We sort all the configurations of text-based clustering on the SEI corpus using mean entropy and mean accuracy, respectively. The results of the top 5 configurations are summarized in Table 4.5.

The top four configurations are the same with either ranking criterion, mean

Table 4.5: The top 5 configurations of text-based clustering on the SEI corpus, sorted in ascending order of mean entropy $\overline{e}$ and in descending order of mean accuracy $\overline{a}$, respectively.

| Rank | $\overline{e}$ | Configuration | $\overline{a}$ | Configuration |
|---|---|---|---|---|
| 1 | 0.9227 | *xm-cnc-tv-300* | 0.6440 | *xm-cnc-tv-300* |
| 2 | 1.0005 | *xm-cnc-chi-300* | 0.6222 | *xm-cnc-chi-300* |
| 3 | 1.0201 | *xm-cnc-tv-500* | 0.6182 | *xm-cnc-tv-500* |
| 4 | 1.0314 | *xm-cnc-chi-500* | 0.6050 | *xm-cnc-chi-500* |
| 5 | 1.0444 | *xm-cnc-mi-500* | 0.6031 | *xm-cnc-tv-200* |

entropy or mean accuracy, and in the same rank order, as shown in Table 4.5.

We also observe that the top five configurations are dominated by $X$-means clustering using CNC document representation, where $X$-means algorithm often returns an optimal number of clusters from 9 to 12. This indicates that $X$-means clustering is better than $K$-means clustering and CNC document representation is better than TFIDF and KEA.

In terms of dimensionality, the top five configurations use a dimensionality between 200 and 500, which indicates feature selection methods can effectively reduce the dimensional space from thousands of features to hundreds.

**Results on the AC Corpus** We sort all the configurations of text-based clustering on the AC corpus using mean entropy and mean accuracy, respectively. The results of the top 5 configurations are presented in Table 4.6.

Table 4.6: The top 5 configurations of text-based clustering on the AC corpus, sorted in ascending order of mean entropy $\overline{e}$ and in descending order of mean accuracy $\overline{a}$, respectively.

| Rank | $\overline{e}$ | Configuration | $\overline{a}$ | Configuration |
|---|---|---|---|---|
| 1 | 0.6200 | *xm-cnc-tv-300* | 0.7502 | *xm-cnc-tv-300* |
| 2 | 0.6267 | *xm-cnc-chi-500* | 0.7426 | *xm-cnc-chi-500* |
| 3 | 0.6389 | *xm-cnc-tv-500* | 0.7347 | *xm-cnc-tv-500* |
| 4 | 0.6795 | *xm-kea-tv-300* | 0.7287 | *xm-kea-tv-300* |
| 5 | 0.6815 | *xm-cnc-chi-300* | 0.7285 | *xm-cnc-chi-300* |

The top five configurations are the same with either ranking criterion, mean entropy or mean accuracy, and in the same rank order, as shown in Table 4.6.

Again we observe that the top configurations are dominated by $X$-means clustering, where an optimal number of clusters from 8 to 12 are returned. This indicates that $X$-means clustering is better than $K$-means clustering.

In terms of dimensionality, the top 5 configurations use a dimensionality between 300 and 500, which indicates feature selection methods can effectively reduce the dimensional space from thousands of features to hundreds.

### Comparison of $K$-means and $X$-means

One of our objectives is to compare the clustering quality of $K$-means and $X$-means algorithms. In order to do this, we pair up clustering configurations such that the only difference between each pair is the clustering method, e.g. *km-cnc-tv-300* vs. *xm-cnc-tv-300*. In this comparison, we use $K = 8$ for the SEI corpus and $K = 9$ for the AC corpus. $X$-means clustering returns the optimal number of clusters (based on Bayesian Information Criterion [58]), which is between 9 and 12 for the SEI corpus, and between 8 and 12 for the AC corpus.

For each pair of clustering configurations, we have a pair of 20 entropy or accuracy values, on which we apply the two-tail paired $t$-test, which generally compares two different methods used for experiments carried in pairs [17]. It is the difference between each pair of measurements which is of interest.

For example, when comparing *km-cnc-tv-300* and *xm-cnc-tv-300*, we have 20 pairs of entropy values, denoted as $e_{km_i}$ and $e_{xm_i}$ ($i = 1, 2, ..., 20$), which are independent observations from the two samples in $K$-means and $X$-means clustering, respectively. Then the differences $d_i = e_{km_i} - e_{xm_i}$ ($i = 1, 2, ..., 20$) will be a sample of size $n$ ($n = 20$) from a population with mean zero. Furthermore, if the populations, from which the above two samples are drawn, are approximately normally distributed, then the differences will also be approximately normally distributed. If the observed average difference is denoted by $\overline{d}$, the standard deviation of the observed differences by $s_d$, and the $t$-test statistic by $t$, then we have the following equations:

$$\overline{d} = \frac{\sum_{i=1}^{n} d_i}{n},$$

$$s_d^2 = \frac{\sum_{i=1}^{n} (d_i - \overline{d})^2}{n-1},$$
$$t = \frac{\overline{d}}{s_d/\sqrt{n}} \ (n = 20). \tag{4.8}$$

The null hypothesis $H_0$ and the alternative hypothesis $H_1$ are given by: 1) $H_0$ : $\overline{d} = 0$ ($K$-Means clustering and $X$-means clustering produce results with the same quality), and 2) $H_1 : \overline{d} > 0$ ($X$-Means clustering is statistically significantly better than $K$-Means clustering). If $H_0$ is true, then the distribution of $t$ will be a $t$-distribution with $n-1$ degrees of freedom, as the estimate $s_d$ is calculated from $n$ differences.

We perform two-tail $t$-tests at the 5% significance level on all 120 pairs[8] of clustering configurations using both entropy and accuracy on the two Web corpora. We observe that in all 480 comparisons the $t$-statistic is greater than $t_{0.05,19}$, which is 2.093 from the $t$-table. Since $t > t_{0.05,19}$ ($P_{value} \leq 0.05$), it is reasonable to reject the null hypothesis $H_0$, i.e., there is a statistically significant difference between the quality values of $K$-means clustering and $X$-means clustering. More precisely, $X$-means statistically significantly outperforms $K$-means in all cases.

In terms of computational performance, we observe that $X$-means clustering is generally faster than $K$-means clustering since the former applies the *KD-trees* data structure for speedup optimization [58]. Critical evaluation of computational performance of both algorithms is not a main goal of this work and consequently it will be one of the future research directions.

### Comparison of Document Representation

We are also interested in learning whether document representation has an impact on the quality of Web page clustering. In order to do this, we pair up clustering configurations such that the only difference between each pair is the document representation method, e.g. *xm-kea-tv-300* vs. *xm-cnc-tv-300*. Since $X$-means clustering is statistically significantly better than $K$-means clustering, we only perform comparisons of document representation methods in $X$-means clustering.

[8] $K$-means vs. $X$-means: 3 (document representations) × 5 (feature selections) × 8 (dimensionalities) = 120 pairs.

We perform $t$-tests at the 5% significance level on all 120 pairs[9] of clustering configurations using both entropy and accuracy measures on the two Web corpora. We observe that the $t$-test results are consistent when either entropy or accuracy is used on both corpora. The $t$-test results of all comparisons are presented in Table 4.7, where $<$ indicates $P_{value} > 0.05$, which means no statistically significant difference at the 5% significance level is found, and $<<$ indicates $P_{value} \leq 0.05$, which means a statistically significant difference at the 5% significance level is found.

Table 4.7: Paired two-tail $t$-tests of document representation methods on the SEI and AC corpora using both entropy and accuracy.

| Method | SEI Corpus | AC Corpus |
|---|---|---|
| TFIDF vs. KEA | << | << |
| TFIDF vs. CNC | << | << |
| KEA vs. CNC | << | < |

Both KEA and CNC are statistically significantly better than TFIDF in $X$-means clustering for both entropy and accuracy evaluation, and CNC is statistically significantly better than KEA on the SEI corpus, as shown in Table 4.7. However, there is no statistically significant difference between KEA and CNC on the AC corpus. This can be explained by the fact that the AC corpus has less narrative text than the SEI corpus and consequently the quality difference between CNC and KEA is smaller on the AC corpus than on the SEI corpus. Overall, CNC is the best document representation method in the Web page clustering task.

### Comparison of Feature Selection

One of our main objectives is to compare the five feature selection methods in the Web page clustering task. Since IG, MI, and CHI are supervised methods, they are treated as upper bounds on the performance of the unsupervised methods DF and TV. In real world clustering applications where topical knowledge is not available, only DF and TV can be used.

In order to perform the comparisons, we pair up clustering configurations such that

[9]TFIDF vs. KEA vs. CNC: $C_3^2 \times 5$ (feature selections) $\times$ 8 (dimensionalities) = 120 pairs.

the only difference between each pair is the feature selection method, e.g. *xm-cnc-tv-300* vs. *xm-cnc-chi-300*. Since $X$-means clustering is statistically significantly better than $K$-means clustering, and CNC is the best document representation method, we only perform comparisons of feature selection methods in $X$-means clustering using the CNC document representation.

We perform $t$-tests at the 5% significance level on all 80 pairs[10] of clustering configurations using both entropy and accuracy measures on the two Web corpora.

**Results on the SEI Corpus** The $t$-test results on the SEI corpus using entropy are presented in Table 4.8, where $<$ or $>$ indicates $P_{value} > 0.05$, which means no statistically significant difference at the 5% significance level is found, and $<<$ or $>>$ indicates $P_{value} \leq 0.05$, which means a statistically significant difference at the 5% significance level is found.

Table 4.8: Paired two-tail $t$-tests of the five feature selection methods on the SEI corpus using entropy.

| Method | TV | IG | MI | CHI |
|---|---|---|---|---|
| DF | << | < | << | << |
| TV | | >> | > | > |
| IG | | | << | << |
| MI | | | | < |

The $t$-test results using entropy can be summarized as TV > CHI > MI >> IG > DF and further grouped as {TV, CHI, MI} >> {DF, IG}[11], as shown in Table 4.8. When accuracy is used, the $t$-test results are similar except that CHI >> MI. Hence, we can conclude that TV is comparable to the supervised methods CHI and MI, and statistically significantly better than the unsupervised method DF. As a result, TV is the feature selection method of choice on the SEI corpus.

**Results on the AC Corpus** We observe that the $t$-test results are consistent when entropy and accuracy are used on the AC corpus, as shown in Table 4.9. Again, $<$ or $>$ indicates $P_{value} > 0.05$, which means no statistically significant difference at the

[10] DF vs. TV vs. IG vs. MI vs. CHI: $C_5^2 \times 8$ (dimensionalities) = 80 pairs.

[11] $>$ or $>>$ is not transitive in the mathematical sense.

5% significance level is found, and << or >> indicates $P_{value} \leq 0.05$, which means a statistically significant difference at the 5% significance level is found.

Table 4.9: Paired two-tail $t$-tests of the five feature selection methods on the AC corpus using both entropy and accuracy.

| Method | TV | IG | MI | CHI |
|---|---|---|---|---|
| DF | << | < | << | << |
| TV | | >> | >> | > |
| IG | | | << | << |
| MI | | | | << |

The $t$-test results can be summarized as TV > CHI >> MI >> IG > DF and further grouped as {TV, CHI} >> {MI} >> {IG, DF}, as shown in Table 4.9. Again, we can conclude that TV is the feature selection method of choice on the AC corpus.

From the results above, we observe that TV is the feature selection method of choice in the Web page clustering task. In most cases, it statistically significantly outperforms MI, IG, and DF. This can be explained by the fact that the topic distribution is imbalanced so there is not enough data for the supervised methods, i.e., IG, MI and CHI, to statistically work well. On the other hand, TV can find features with higher variance across documents that have more power to discriminate documents. However, comparing these feature selection methods on more Web sites is highly desired.

### Comparison of Dimensionalities

We perform text-based clustering in both high and low dimensional space in order to find out if feature selection methods can significantly reduce the dimensionality in text-based clustering. We want to find out what is the proper range of the number of features (instead of a fixed number of features) for each feature selection method such that when these features are used the best clustering quality can be achieved.

Since $X$-means clustering is statistically significantly better than $K$-means clustering and CNC is the best document representation method, we only perform comparisons of dimensionality in $X$-means clustering using CNC document representation. Each feature selection method is separately used to rank the CNC phrases and

the top $n$ ($n \in \{50, 100, 200, 300, 500, 1000, 2000, 3000\}$) phrases are used as features. For each feature selection method, we simply rank the quality of clustering using 8 different dimensionalities according to mean entropy or mean accuracy.

We observe that the results of dimensionality comparisons are consistent when entropy and accuracy are used on both corpora. For each feature selection method on the two corpora, we use both entropy and accuracy to choose the best dimensionality, which is presented in Table 4.10.

Table 4.10: Best dimensionality for each feature selection method in CNC-based $X$-means clustering on the two corpora using both entropy and accuracy.

| Method | DF | TV | IG | MI | CHI |
|---|---|---|---|---|---|
| SEI | 200 | 300 | 100 | 500 | 300 |
| AC | 300 | 300 | 200 | 500 | 500 |

DF and IG tend to suggest a smaller dimensionality than TV, CHI, and MI, as shown in Table 4.10. Generally speaking, a dimensionality around 300 seems to be a reasonable choice for most feature selection methods. This reduces the high dimensional space of text-based clustering, which is often as high as 3000, by an order of magnitude.

### 4.4.3 Link-based Clustering

In addition to the conventional text-based clustering, we also investigate the quality of Web page clustering using link-based features, i.e., the incoming and outgoing links. For link-based clustering, we are interested in finding out whether there is a statistically significant difference between the quality of the following clustering options:

- Clustering methods: $K$-means vs. $X$-means. Same as text-based clustering, we set $K$ equal to the number of topics in each Web corpus for $K$-means clustering, i.e., 8 for the SEI corpus and 9 for the AC corpus. For $X$-means clustering, we let $X$-means determine the optimal number of clusters within the range of [1, 20].

- Document representation: co-citation vs. bibliographic coupling. Web pages whose depth value is less than a pre-defined threshold form the feature space. We separately apply co-citation and bibliographic coupling to represent each Web page using a link-based vector $V$, denoted by $\{w_1, w_2, ..., w_k\}$ $(k \leq N)$, where $w_i$ equals to 1 if page $p_i$ (represented by its URL $u_i$) appears as an incoming or outgoing link of the target page, and 0 otherwise.

- Dimensionality: we perform clustering using the top $k$ pages of the feature space, which consists of pages whose depth value is less than a pre-defined depth value. The eight different dimensionalities we choose are: 30, 50, 100, 150, 200, 300, 500, and all.

We enumerate all configurations of the above options to evaluate link-based clustering. This leads to a total of 2 (algorithms) $\times$ 2 (document representations) $\times$ 8 (dimensionalities) $=$ 32 clustering configurations for each Web corpus. We denote each clustering configuration in the order of *clustering method*, *document representation*, and *dimensionality*. For instance, the configuration of *K-means clustering using co-citation with the top 200 pages* will be denoted as *km-in-200*, and the configuration of *X-means clustering using bibliographic coupling with the top 300 pages* will be denoted as *xm-out-300*.

For each configuration, the clustering is repeated for 20 times using the 20 randomly chosen seeds, which produces a list of 20 entropy and 20 accuracy values. The mean entropy (denoted as $\overline{e}$), or the mean accuracy (denoted as $\overline{a}$) over all 20 runs is taken as the quality of this particular clustering.

### Comparison of All Configurations

We aim to find out which link-based clustering configuration can lead to the best clustering quality. For each Web corpus, we sort all the configurations in ascending order of mean entropy (the lower, the better) and in descending order of mean accuracy (the higher, the better), respectively.

**Results on the SEI Corpus** We sort all the configurations of link-based clustering on the SEI corpus using mean entropy and mean accuracy, respectively. The results of the top 5 configurations are summarized in Table 4.11.

Table 4.11: The top 5 configurations of link-based clustering on the SEI corpus, sorted in ascending order of mean entropy $\overline{e}$ and in descending order of mean accuracy $\overline{a}$, respectively.

| Rank | $\overline{e}$ | Configuration | $\overline{a}$ | Configuration |
|---|---|---|---|---|
| 1 | 0.6296 | *xm-out-300* | 0.7605 | *xm-out-300* |
| 2 | 0.6450 | *xm-out-200* | 0.7579 | *xm-out-200* |
| 3 | 0.8421 | *xm-out-150* | 0.6855 | *xm-out-150* |
| 4 | 0.8628 | *xm-out-100* | 0.6628 | *xm-out-100* |
| 5 | 0.9243 | *xm-out-500* | 0.6341 | *xm-out-500* |

The top five configurations are the same with either ranking criterion, mean entropy or mean accuracy, and in the same rank order, as shown in Table 4.11.

We also observe that the top five configurations are dominated by $X$-means clustering using bibliographic coupling. This indicates that $X$-means clustering is better than $K$-means clustering and that bibliographic coupling is better than co-citation. The best clustering configuration using co-citation is *xm-in-200*, which achieves a mean entropy of 1.2964 and a mean accuracy of 0.4983.

In terms of dimensionality, the top five configurations use a dimensionality between 150 and 500, which indicates that link-based clustering only requires a few hundreds of pages as features.

**Results on the AC Corpus** We sort all the configurations of link-based clustering on the AC corpus using mean entropy and mean accuracy, respectively. The results of the top 5 configurations are summarized in Table 4.12.

Table 4.12: The top 5 configurations of link-based clustering on the AC corpus, sorted in ascending order of mean entropy $\overline{e}$ and in descending order of mean accuracy $\overline{a}$, respectively.

| Rank | $\overline{e}$ | Configuration | $\overline{a}$ | Configuration |
|---|---|---|---|---|
| 1 | 1.8082 | *xm-out-200* | 0.3589 | *xm-out-200* |
| 2 | 1.8227 | *xm-out-150* | 0.3337 | *xm-out-150* |
| 3 | 1.9028 | *xm-out-100* | 0.3048 | *xm-out-100* |
| 4 | 1.9614 | *xm-out-300* | 0.2935 | *xm-out-300* |
| 5 | 1.9982 | *xm-out-50* | 0.2682 | *xm-out-50* |

Same as link-based clustering on the SEI corpus, the ordering of the top five configurations using mean entropy for ranking is exactly the same as those using mean accuracy for ranking, as shown in Table 4.12. Moreover, the top five configurations are again dominated by $X$-means clustering using bibliographic coupling. This is a strong indication that $X$-means is better than $K$-means and that bibliographic coupling is better than co-citation in the Web page clustering task.

The best clustering configuration using co-citation is *xm-in-100*, which achieves a mean entropy of 2.2964 and a mean accuracy of 0.2383.

In terms of dimensionality, the top 5 configurations use a dimensionality between 50 and 300, which indicates that link-based clustering only requires a small portion of all pages as features.

### Comparison of $K$-means and $X$-means

Same as the comparison of $K$-means and $X$-means in text-based clustering, we compare the two methods in link-based clustering. Again, we pair up clustering configurations such that the only difference between each pair is the clustering method, e.g. *km-out-300* vs. *xm-out-300*.

We perform $t$-tests at the 5% significance level on all 16 pairs[12] of clustering configurations using both entropy and accuracy on the two Web corpora. We observe that $X$-means statistically significantly outperforms $K$-means in all 64 comparisons.

### Comparison of Document Representation

We are interested in learning whether there is a statistically significant difference between link-based clustering using co-citation and bibliographic coupling. In order to do this, we pair up clustering configurations such that the only difference between each pair is the document representation method, e.g. *xm-in-300* vs. *xm-out-300*. Since $X$-means clustering is statistically significantly better than $K$-means clustering, we only perform comparisons of document representation methods in $X$-means clustering.

We perform $t$-tests at the 5% significance level on all 8 pairs[13] of clustering configurations using both entropy and accuracy measures on the two Web corpora. We

[12] $K$-means vs. $X$-means: 2 (document representations) $\times$ 8 (dimensionalities) = 16 pairs.
[13] Co-citation vs. Bibliographic Coupling: one pair for each of 8 dimensionalities.

observe that in all 32 comparisons bibliographic coupling is statistically significantly better than co-citation.

### Comparison of Dimensionalities

We perform link-based clustering in both high and low dimensional space in order to find out what is the proper range of the number of links (instead of a fixed number of links) such that when these links are used the best clustering quality can be achieved.

Since $X$-means clustering is statistically significantly better than $K$-means clustering and bibliographic coupling statistically significantly outperforms co-citation, we only perform comparisons of dimensionality in $X$-means clustering using bibliographic coupling.

For the two Web corpora, we use the top $k$ ($k \in \{30, 50, 100, 150, 200, 300, 500, all\}$) pages as features. We rank the quality of clustering using 8 different dimensionalities according to mean entropy or mean accuracy.

We observe that the results of dimensionality comparisons are consistent when entropy and accuracy are used on both corpora. The best two dimensionalities for the SEI corpus are 300 and 200, as shown in Table 4.11. For the the AC corpus, the best two dimensionalities are 200 and 150. Generally speaking, a dimensionality of around 200 seems to be a reasonable choice.

## 4.4.4 Coupled Clustering

We have presented the results of text-based and link-based clustering, respectively. It is of interest to see whether there is any quality difference between these two approaches and whether combining them can gain more improvement of clustering quality.

### Comparison of Text- and Link-based Clustering

We aim to find out which of the two methods, text-based or link-based clustering, can achieve higher quality on both Web corpora. We perform $t$-tests at the 5% significance level on pairs of the top 5 text- and link-based clustering configurations using both entropy and accuracy, where each pair is the $i^{th}$ best text-based configuration versus the $i^{th}$ ($1 \leq i \leq 5$) best link-based configuration.

We compare the best five text-based clustering configurations with the best five link-based clustering configurations on both Web corpora (Tables 4.5, 4.11 for the SEI corpus, and Tables 4.6, 4.12 for the AC corpus), as shown in Table 4.13. We observe that the results using entropy are the same as those using accuracy.

Table 4.13: Paired two-tail $t$-tests of the best five text-based clustering versus the best five link-based clustering on both Web corpora.

| Rank | SEI Corpus | AC Corpus |
|---|---|---|
| 1 | *xm-cnc-tv-300 << xm-out-300* | *xm-cnc-tv-300 >> xm-out-200* |
| 2 | *xm-cnc-chi-300 << xm-out-200* | *xm-cnc-chi-500 >> xm-out-150* |
| 3 | *xm-cnc-tv-500 << xm-out-150* | *xm-cnc-tv-500 >> xm-out-100* |
| 4 | *xm-cnc-chi-500 << xm-out-100* | *xm-kea-tv-300 >> xm-out-300* |
| 5 | *xm-cnc-mi-500 << xm-out-500* | *xm-cnc-chi-300 >> xm-out-50* |

Link-based clustering is statistically significantly better than text-based clustering on the SEI corpus, as shown in Table 4.13. However, it is the opposite on the AC corpus, i.e., text-based clustering statistically significantly outperforms link-based clustering. We hypothesize that the effectiveness of link-based clustering depends on the richness of linkage information. The SEI Web site is more of an organically grown Web site with rich cross links, whereas the AC Web site is a corporate hierarchical Web site. To confirm this claim, we take a further look at the outgoing links of both corpora. There is an average of 15.5 outgoing links for a SEI Web page, but only 6.8 outgoing links for an AC Web page. This indicates that only when the linkage information is rich can the link-based clustering achieve high quality.

### Coupled Text- and Link-based Clustering

We have seen that link-based clustering can be very effective such as on the SEI corpus. We aim to find a measure to detect how rich is the linkage information available in a Web site and how heavily should the linkage information be used to complement text-based clustering.

In order to achieve this, we combine text- and link-based features to perform clustering. Each document will be represented by a single vector, which consists of two sub-vectors, one with text-based features, and the other with link-based features.

Based on the evaluation results of text- and link-based clustering, we choose the best text- and link-based clustering configurations, i.e., *xm-cnc-tv-300* and *xm-out-300* on the SEI corpus, and *xm-cnc-tv-300* and *xm-out-200* on the AC corpus, respectively.

Let $V$ denote a document vector, $V_{text}$ the text-based sub-vector, and $V_{link}$ the link-based sub-vector, respectively. We combine text- and link-based sub-vectors using a linear model, i.e., $V = \{\lambda \cdot V_{text}, (1-\lambda) \cdot V_{link}\}$ ($\lambda \in [0,1]$). The key is to determine $\lambda$, i.e., how much weight we should give to each of the two sub-vectors.

Finding the best $\lambda$ is a one-dimension optimal search problem. To simplify the problem, we choose $\lambda$ from 0 to 1 in increasing steps of 0.1 to perform $X$-means clustering using the combined features. Entropy and accuracy values are calculated the same as before. The mean entropy and accuracy values for each coupled clustering (a different $\lambda$) are summarized in Table 4.14.

Table 4.14: Mean entropy and mean accuracy values for coupled text- and link-based clustering on the SEI and AC corpora.

| | SEI Corpus | | AC Corpus | |
|---|---|---|---|---|
| $\lambda$ | $\overline{e}$ | $\overline{a}$ | $\overline{e}$ | $\overline{a}$ |
| 0.0 | 0.6296 | 0.7605 | 1.8082 | 0.3589 |
| 0.1 | 0.6268 | 0.7641 | 1.7235 | 0.4386 |
| 0.2 | 0.6209 | 0.7687 | 1.5923 | 0.4716 |
| 0.3 | 0.6152 | 0.7696 | 1.5085 | 0.5186 |
| 0.4 | 0.6098 | 0.7757 | 1.4393 | 0.5537 |
| 0.5 | 0.6342 | 0.7586 | 1.3187 | 0.5969 |
| 0.6 | 0.6937 | 0.7239 | 1.2234 | 0.6573 |
| 0.7 | 0.7561 | 0.7015 | 0.9127 | 0.6858 |
| 0.8 | 0.8123 | 0.6890 | 0.7239 | 0.7012 |
| 0.9 | 0.8639 | 0.6803 | 0.6456 | 0.7149 |
| 1.0 | 0.9227 | 0.6440 | 0.6200 | 0.7502 |

The best $\lambda$ for the coupled clustering on the SEI corpus is 0.4, as shown in Table 4.14. This means combining text- and link-based features achieves better clustering quality than using either text- or link-based features alone. However, on the AC corpus, incorporating link-based features always decreases the clustering quality, which means link-based features are useless.

The above results indicate that text is more consistently reliable for Web page

clustering than link knowledge. For some Web sites (e.g. SEI), linkage information is helpful. For other sites (e.g. AC), it might be harmful (at least no benefit is gained). The link structure of a Web site is more like a "tree" with back links[14] and cross links[15] [70]. It is different from the Web graph, where link structure has already been shown to be useful in various Web-based applications.

We observe that the average number of cross links (or the ratio of cross links in outgoing links) of a Web page is an indicator of whether the linkage information should be incorporated into clustering. If the cross link information is rich (e.g. more than 50% of outgoing links are cross links), then giving higher weight to link-based features will achieve better clustering quality.

This makes sense because intuitively commercial Web sites (e.g. AC) are more likely to be designed and constructed by a person or a team using specific tools. Consequently they are more likely to be hierarchical (a tree with branches and some back links). In contrast, academic Web sites (e.g. SEI) are often built and connected by many individuals. Moreover, they tend to grow organically and have more cross links between nodes.

Precisely measuring the richness of cross link information is a future research direction.

## 4.5 Summary

In this thesis, we investigate $K$-means and $X$-means clustering using both text- and link-based features. In text-based clustering, we study document representation methods and feature selection methods for dimensionality reduction. In the link-based clustering, we study co-citation and bibliographic coupling. We evaluate the clustering quality using both entropy and accuracy, which are found to be consistent. Our main contribution consists of the following findings:

- $X$-means algorithm is statistically significantly better than $K$-means algorithm in the Web page clustering task.

[14] A back link is a hyperlink from a lower depth level Web page pointing to a higher depth level Web page, where levels are determined in the breadth-first site traversal.

[15] A cross link is a hyperlink between two Web pages, which are at the same depth level of the breadth-first site traversal.

- CNC is the best text-based document representation method.
- Term Variance is the best text-based feature selection method, which can reduce the dimensionality by an order of magnitude if CNC is used.
- Bibliographic coupling is statistically significantly better than co-citation in the link-based clustering.
- Combining text- and link-based features can improve the quality of clustering using either type of features alone if the cross link information is rich.

# Chapter 5

# Cluster Summarization

In the previous chapter, we present how to conduct coupled text and link-based clustering to obtain significant topic groups of a given Web site. In this chapter, we discuss how to separately summarize each individual cluster. We also present summaries of test Web sites, experimental methodology, and evaluation results.

Automatic summarization of an individual cluster is a multi-step process adapted from our previous work [85]. It consists of the following steps.

1. First, key phrases are extracted from the narrative text of Web pages. The key phrase extraction tool we use in this thesis is the C-value/NC-value (CNC) [22] method, which has been found to outperform alternative key phrase extraction methods in Chapter 3.

2. Second, key sentences are extracted from the narrative text of all Web pages. In this thesis, we propose a classification method where linguistic and lexical features are used to build a classifier that can be used to classify a sentence into key-sentence or non-key-sentence automatically.

3. Third, a short summary is generated for each cluster. The cluster summary consists of the top $n_1$ key phrases and the top $n_2$ key sentences. The parameters $n_1$ and $n_2$ are heuristically determined by both the informativeness of the key phrases and the key sentences and the size of the cluster summaries.

## 5.1 Key Sentence Extraction

Traditionally, in an extraction-based summarization system, once the key phrases are identified, the most significant sentences can be retrieved based on the density of key phrases present in them [11]. The significance of a sentence is often measured by calculating an importance value, which is the maximum of weights of all *word clusters*

within the sentence. A word cluster is defined as a sequence of words which starts and ends with a key phrase and at most 2 non-key-phrases must separate any two neighboring key phrases [8]. The weight of a word cluster is computed by adding the weights of all key phrases within the word cluster, and dividing this sum by the total number of key phrases [85]. The maximum weight of word clusters is taken as the *sentence weight*. All sentences in narrative text paragraphs are ranked by *sentence weight* and the top sentences are the key sentences to be included in the summary.

Intuitively, whether a sentence is a key sentence or not is mainly determined by its coherence and topicality (relatedness to the main topic of the target Web site). The coherence is more or less reflected by the part-of-speech patterns, which have proved to be effective in several Web-based applications such as query ambiguity reduction [1] and question answering [59]. The topicality has strong connection with features such as the depth level of the Web page where a sentence appears, as well as its weight calculated above. We hypothesize that these linguistic and lexical features contain sufficient information to determine whether a sentence is a key sentence or not. We apply the classification method to address this problem.

### 5.1.1 Key Sentence Classification

In order to build *KeySentence*, a classifier that is able to classify a sentence as key-sentence or non-key-sentence, a data set is needed. In our previous approach [85], we created a collection of 3242 paragraphs for learning the NARRATIVE classifier, which classifies a text paragraph into narrative or non-narrative. From the paragraphs labelled narrative, we randomly select 1328 sentences. Then, the part-of-speech tags for all words in these sentences are computed using a rule-based part-of-speech tagger [5]. A total of 32 part-of-speech tags are found and summarized in Table 5.1.

Each part-of-speech tag is quantified by its frequency of occurrence in a paragraph. Let $n_i$ (i = 1, 2, ..., 32) be the number of times the $i^{th}$ tag appears in the paragraph. Then $P_i$, the fraction of the total number of all 32 tags (i.e., words) that $n_i$ represents, is represented by Equation 5.1.

$$P_i = \frac{n_i}{\sum_{i=1}^{32} n_i} \tag{5.1}$$

Eight more attributes are added to this feature set for building the KeySentence classifier. All the 40 features and their meanings for a sentence are summarized

Table 5.1: A list of 32 part-of-speech tags used in part-of-speech tagging.

| Tag | Meaning & Example | Tag | Meaning & Example |
|---|---|---|---|
| CC | conjunction (and, or) | RBR | adverb, comparative (faster) |
| CD | number (four, fourth) | RBS | adverb, superlative (fastest) |
| DT | determiner, general (a, the) | RB | adverb, general (fast) |
| EX | existential (there) | SYM | symbol or formula (US$500) |
| FW | foreign word (ante, de) | TO | infinitive marker (to) |
| IN | preposition (on, of) | UH | interjection (oh, yes, no) |
| JJR | adjective, comparative (lower) | VBD | verb, past tense (went) |
| JJS | adjective, superlative (lowest) | VBG | verb, -ing (going) |
| JJ | adjective, general (near) | VBN | verb, past participle (gone) |
| MD | modal auxiliary (might, will) | VBP | verb, (am, are) |
| NNPS | noun, proper plural (Americas) | VBZ | verb, -s (goes, is) |
| NNP | noun, proper singular (America) | VB | verb, base (go, be) |
| NNS | noun, common plural (cars) | WDT | det, wh- (what, which) |
| NN | noun, common singular (car) | WP$ | pronoun, possessive (whose) |
| PRP$ | pronoun, possessive (my, his) | WP | pronoun (who) |
| PRP | pronoun, personal (I, he) | WRB | adv, wh- (when, where, why) |

in Table 5.2. Next, each sentence is manually labelled as key-sentence or non-key-sentence. The criterion to determine if a sentence is a key sentence or not is that a key sentence must provide important topical information about the Web site. Thus, we obtain a data set of 1328 sentences for building the KeySentence classifier.

We use the machine learning tool C5.0 to perform key sentence classification because of its conceptual simplicity and generally good quality. The 10-fold cross-validation shows a mean error rate of 9.5%, which is summarized in Table 5.3.

The decision tree generated by the C5.0 program and its evaluation are presented in Figure 5.1. The size is 11, i.e., the tree has 11 leaf nodes. Among the total 1328 cases, 116 cases are misclassified, leading to an error of 8.7%. In the decision tree, about 29.5% of cases are following this rule: if the percentage of general determiners is not greater than 3.125%, and the percentage of proper singular nouns is greater than 12.1212%, then the sentence is not a key sentence.

A total of seven features play an important role in this decision tree, i.e., $P_{DT}$, $P_{NNP}$, $P_{CC}$, $distance_d$, $length$, $depth$, and $frequency$, as shown in Figure 5.1. This demonstrates that our assumption is true, i.e., linguistic and lexical features have a

Table 5.2: A list of 40 features used in learning the KeySentence classifier.

| **Feature** | **Meaning** |
|---|---|
| $P_i$ | percentage of each of 32 part-of-speech tags |
| *length* | number of words in a sentence |
| *depth* | depth level of the Web page |
| $distance_d$ | number of words from the beginning of document |
| $distance_p$ | number of words from the beginning of paragraph |
| *phrase* | number of key phrases |
| *frequency* | sum of number of occurrences of all key phrases |
| *cluster* | number of word clusters |
| *weight* | sentence weight |

Table 5.3: Cross-validation of the KeySentence classifier.

| Fold | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Size | 11 | 12 | 12 | 13 | 10 | 11 | 13 | 12 | 11 | 10 | 11.5 |
| Error(%) | 9.5 | 10.2 | 10.3 | 9.7 | 8.9 | 11.2 | 9.6 | 8.3 | 8.5 | 9.0 | 9.5 |

huge impact in determining the coherence and topical significance of a sentence.

We apply the KeySentence classifier on all sentences from the narrative paragraphs of an individual cluster. The key sentences are then sorted according to the descending order of their significance weight. Often the top 5 key sentences are used to form the cluster summary, together with the key phrases.

## 5.2 Cluster Summary Formation

Once the key phrases and key sentences for a given cluster are extracted, it is straightforward to generate a short cluster summary, which consists of the top $n_1$ key phrases and the top $n_2$ key sentences. The parameters $n_1$ and $n_2$ are heuristically determined by both the informativeness of the key phrases and the key sentences and the size of the cluster summaries. In this thesis, we set $n_1 = 5$ and $n_2 = 5$.

$P_{DT} \leq 0.03125$ :
    $P_{NNP} \leq 0.121212 : y$ (68/12)
    $P_{NNP} > 0.121212 : n$ (392/16)
$P_{DT} > 0.03125$ :
    $P_{NNP} > 0.538462$ :
        $distance_d \leq 0.127796 : y$ (16)
        $distance_d > 0.127796 : n$ (76/12)
    $P_{NNP} \leq 0.538642$ :
        $length \leq 11$ :
            $distance_d \leq 0.308824 : y$ (16/4)
            $distance_d > 0.308824 : n$ (20)
        $length > 11$ :
            $P_{CC} > 0.045455 : y$ (256/12)
            $P_{CC} \leq 0.045455$ :
                $P_{DT} > 0.125 : y$ (252/20)
                $P_{DT} \leq 0.125$
                    $depth > 3 : y$ (172/36)
                    $depth \leq 3$
                        $frequency \leq 3 : n$ (40/4)
                        $frequency > 3 : y$ (20)

| $(y)$ | $(n)$ | $\leftarrow$ classified as |
|---|---|---|
| 716 | 32 | $(y)$: class *key-sentence* |
| 84 | 496 | $(n)$: class *non-key-sentence* |

Figure 5.1: Decision tree of the KeySentence classifier.

## 5.3 Experiments and Evaluation

In this section, we show how summaries of test Web sites are generated, describe the methodology of our user study, and present the evaluation results.

### 5.3.1 Summaries of Test Web Sites

In this thesis, we choose six Web sites from the DMOZ directory. The URLs are listed in Table 5.4.

The six sites have been widely tested in our previous summarization research [85, 83, 86]. The first three sites are academic Web sites regarding software engineering

Table 5.4: URLs of the six test Web sites chosen from the DMOZ directory.

| Index | Site | URL |
|---|---|---|
| 1 | SEI | http://www.sei.cmu.edu |
| 2 | AIAI | http://www.aiai.ed.ac.uk |
| 3 | AI | http://www.ai.uga.edu |
| 4 | AC | http://www.aircanada.ca |
| 5 | Nortel | http://www.nortel.com |
| 6 | Oracle | http://www.oracle.com |

and artificial intelligence, while the last three sites are commercial airlines and network product/service providers.

For each test Web site, we apply the clustering approach to obtain the main topics. Some clusters have too few documents for summarization. Consequently, we sort all resulting clusters according to the number of documents in each cluster. The top five clusters are summarized using the CNC key phrase extraction method and the KeySentence classifier introduced in 5.1. Each cluster summary consists of 5 key phrases and 5 key sentences. These five short summaries represent the most important topics covered in the given Web site, and they will be compared with a single long summary, which is generated using the CNC-based summarization system and consists of 25 key phrases and 25 key sentences.

A full list of all short cluster summaries and single long summaries are presented in Appendix C.

### 5.3.2 Evaluation Methodology

We conduct a user study to evaluate the extraction-based summarization system (see Section 3.3.2). Similarly, we aim to conduct another user study to compare the long single summary and the five short cluster summaries of a given Web site.

A size of 20 subjects is sufficient for our study. Each subject was asked to review all the six Web sites. This means that for each type of summary, we have a sample size of 120 with replication. Participants are graduate students in computer science with strong reading comprehension and significant Web browsing experience.

Human subjects are provided with instructions on how to conduct the user study,

including 1) a summary rating example, which is designed to help them be familiar with the evaluation process; and 2) a short survey, which is designed to get feedback on improving our summarization systems. The summary rating example and the short survey are presented in Appendix B.

For each Web site, human subjects are asked to execute the following steps:

1. Browse the Web site and subjectively define *the most essential topic*, which is defined as *the entity behind the Web site and its main activity.* The most essential topic serves as a representation of the core contents of the target Web site. For example, the most essential topic for the SEI Web site could be defined as "Software Engineering Institute at CMU for improvement of software engineering management and practice".

2. Read two types of automatically generated summaries: a) a single long summary, which is obtained by summarizing an entire Web site directly, consisting of 25 key phrases and 25 key sentences; and b) the top five cluster summaries, which are obtained by summarizing the five largest clusters after clustering Web pages in the given site, each consisting of 5 key phrases and 5 key sentences, and then decide their informativeness.

3. Rate each summary element (key phrase or key sentence) based on the extent to which it is related to the most essential topic, using a 1-to-5 scale (1 = not related, 2 = poorly related, 3 = fairly related, 4 = well related, and 5 = strongly related).

4. Complete a short survey, which is designed to get feedback on improving our summarization systems.

### 5.3.3 Summary Evaluation

In this subsection, we explain how to measure the quality of the single long summary and the top five short cluster summaries, respectively. We also present the statistical analysis of rating data collected in the user study. Our main objective is to investigate which type of summary captures better the topics and contents covered in a Web site.

### Evaluation Measures

For each type of summary, we have a sample size of 120 (6 Web sites, and 20 subjects for each site) with replication. Let $n_1$, $n_2$, $n_3$, $n_4$, and $n_5$ be the number of summary elements (key phrases or key sentences) that receive a score of 1, 2, 3, 4, and 5, respectively. Hence for each summary, $\sum_{i=1}^{5} n_i$ will be 25 for either key phrases or key sentences.

We are interested in the distribution of users' rating scores. The percentage of numbers of times each score is given for both types of summaries are shown in Table 5.5. As we can see, the individual cluster summaries have a higher concentration of scores of 4 and 5 over the single summary of the overall Web site and close to each other for the score of 3.

Table 5.5: Distribution of users' rating scores for two types of summaries.

| | Key Phrases | | | | | Key Sentences | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Score | 5 | 4 | 3 | 2 | 1 | 5 | 4 | 3 | 2 | 1 |
| Single (%) | 16.9 | 25.9 | 26.9 | 17.4 | 13.0 | 12.2 | 25.9 | 28.0 | 23.8 | 10.1 |
| Cluster (%) | 24.8 | 32.9 | 23.7 | 12.8 | 5.9 | 19.2 | 30.3 | 27.5 | 15.4 | 7.6 |

We aim to formally evaluate and compare the five key phrase extraction methods by an analysis of both *acceptable percentage* and *quality value* (see Section 3.3.3), which are both calculated based on the rating data obtained in the user study.

**Acceptable Percentage** We are interested in the extent to which the summary elements are acceptable to human readers. In our thesis, acceptable key phrases and key sentences are those that receive a score of 3, 4, or 5. The percentage, $P$, is formally defined in Equation 3.6.

For each type of summary (either the single long summary or the top five cluster summaries), let $P_{kp}$ be the acceptable percentages of key phrases and $P_{ks}$ be the acceptable percentage of key sentences, respectively. Then the final score of a summary, denoted as $P_s$, is a linear combination of $P_{kp}$ and $P_{ks}$, i.e., $P_s = \lambda \cdot P_{kp} + (1 - \lambda) \cdot P_{ks}$, $\lambda \in (0, 1)$. The $\lambda$ is empirically set to 0.5 based on users' preference in a simple survey conducted in the key phrase extraction research (see Section 3.3.3).

**Quality Value** In addition to the acceptable percentage measure, we also aim to compare the two types of summaries using the quality values of summary elements. The average quality, $Q$, of 25 key phrases or key sentences in a summary is the average score achieved by the elements of the summary, formally defined in Equation 3.7.

For each type of summary (either the single long summary or the top five cluster summaries), let $Q_{kp}$ be the average quality of key phrases and $Q_{ks}$ be the acceptable percentage of key sentences, respectively. Then the final quality of a summary, denoted as $Q_s$, is a linear combination of $Q_{kp}$ and $Q_{ks}$, i.e., $Q_s = \lambda \cdot Q_{kp} + (1 - \lambda) \cdot Q_{ks}$, $\lambda \in (0, 1)$. Again, the $\lambda$ is empirically set to 0.5.

### Evaluation Results

The results of acceptable percentage values for both types of summaries of the six test Web sites are presented in Table 5.6.

Table 5.6: Summary of acceptable percentage for both types of summaries of the six test Web sites.

| | Single summary | | | Cluster summaries | | |
|---|---|---|---|---|---|---|
| Site | $P_{kp}$ | $P_{ks}$ | $P_s$ | $P_{kp}$ | $P_{ks}$ | $P_s$ |
| SEI | 0.82 | 0.87 | 0.85 | 0.89 | 0.95 | 0.92 |
| AIAI | 0.63 | 0.61 | 0.62 | 0.75 | 0.68 | 0.71 |
| AI | 0.70 | 0.81 | 0.75 | 0.82 | 0.83 | 0.82 |
| AC | 0.52 | 0.54 | 0.53 | 0.68 | 0.51 | 0.60 |
| Nortel | 0.71 | 0.58 | 0.65 | 0.84 | 0.82 | 0.83 |
| Oracle | 0.79 | 0.56 | 0.68 | 0.90 | 0.83 | 0.87 |
| Average | 0.70 | 0.66 | 0.68 | 0.81 | 0.77 | 0.79 |

The acceptable percentage of the summaries of top five clusters is higher than that of the single long summary for each test Web site, as shown in Table 5.6. We apply the One-Way Fully Repeated Measures ANOVA on the acceptable percentage data and a statistically significant difference between the two types of summaries ($P_{value} = 4.92065E^{-06}$) is found at the 5% significance level. This indicates that the top five cluster summaries statistically significantly outperform the single long summary in capturing the essential topics and main contents covered in a Web site.

We are generous in the sense that we take summary elements of score 3 as acceptable. If we are stricter and take only summary elements of scores 4 and 5 as acceptable, then the acceptable percentage difference between two types of summaries will be even bigger, as can be seen in Table 5.5.

The results of average quality values for both types of summaries of the six test Web sites are presented in Table 5.7.

Table 5.7: Summary of average quality for both types of summaries of the six test Web sites.

| | Single summary | | | Cluster summaries | | |
|---|---|---|---|---|---|---|
| Site | $Q_{kp}$ | $Q_{ks}$ | $Q_s$ | $Q_{kp}$ | $Q_{ks}$ | $Q_s$ |
| SEI | 3.36 | 3.60 | 3.48 | 3.78 | 4.07 | 3.92 |
| AIAI | 2.97 | 2.93 | 2.95 | 3.31 | 3.07 | 3.19 |
| AI | 3.12 | 3.28 | 3.20 | 3.62 | 3.45 | 3.54 |
| AC | 2.81 | 2.79 | 2.80 | 3.34 | 2.77 | 3.05 |
| Nortel | 3.19 | 2.88 | 3.03 | 3.52 | 3.39 | 3.45 |
| Oracle | 3.52 | 2.91 | 3.21 | 3.91 | 3.57 | 3.74 |
| Average | 3.16 | 3.06 | 3.11 | 3.58 | 3.39 | 3.48 |

The average quality of the summaries of the top five clusters is higher than that of the single long summary for each test Web site, as shown in Table 5.7. We apply the One-Way Fully Repeated Measures ANOVA on the quality value data and a statistically significant difference between the two types of summaries ($P_{value} = 7.81804E^{-08}$) is found at the 5% significance level.

The acceptable percentage measure and the quality value measure lead to the same evaluation results, i.e., the top five clusters summaries are statistically significantly better than the single long summary. This can be explained by the fact that the acceptable percentage and the average quality are intrinsically related as they are both based on users' ratings. The only difference is that the former gives equal weight to (a summation of) the number of summary elements with scores 3, 4, and 5, while the latter gives different weight to summary elements with different scores (number of such elements times the score they receive).

### Analysis of a Survey

We have demonstrated that the top five cluster summaries are statistically significantly better than the single long summary in the multi-topic Web site summarization task. Moreover, we aim to present the results of the short survey completed by the study participants.

- Survey Question 1. *Overall, which type of summary do you like better, the single long summary or the top five short cluster summaries?*

  A total of 17 subjects vote for the top five short clusters summaries and the other three favor the single long summary, which has been shown to be comparable to human authored summaries [85]. People seem to prefer the cluster summaries because of the following reasons:

  - The cluster summaries often reveal the main topics on a web site.
  - The cluster summaries are shorter and more representative of the main topics.
  - The cluster summaries are more comprehensive and comprehensible.

- Survey Question 2. *If the answer to Question 1 is the latter, how many cluster summaries do you prefer?*

  The 17 subjects, who like the cluster summaries better, prefer the number of cluster summaries to be between 3 and 6. The average of their preferred numbers is 4.9, which is very close to the number we choose, i.e., 5. The standard deviation is 0.8.

- Survey Question 3. *For the single long summary, what is the ideal number of key phrases in your opinion? And key sentences?*

  The ideal number of key phrases given by human subjects varies from 5 to 15, with an average of 9.8 and a standard deviation of 2.7. The ideal number of key sentences varies from 5 to 10, with an average of 7.1 and a standard deviation of 1.6. This indicates that 10 and 7 might be reasonable numbers for key phrases and key sentences, respectively.

- Survey Question 4. *For each cluster summary, do the same as in Question 3.*

  The ideal number of key phrases in each cluster summary is suggested to be between 3 and 6, leading to an average of 4.7 and a standard deviation of 0.9. The ideal number of key sentences suggested varies from 2 to 5, leading to an average of 3.3 and a standard deviation of 1.2. This indicates that for each cluster summary, 5 key phrases and 3 key sentences are acceptable.

- Survey Question 5. *The summary consists of two parts, key phrases and key sentences. Which part do you think is more important? Suppose there are 10 points of importance for the whole summary. How many points should be assigned to each part (e.g. 4 to key phrases and 6 to key sentences), respectively?*

  The points given by subjects vary from 3:7 (meaning 3 for key phrases and 7 for key sentences) to 6:4, leading to an average of 5:5. The standard deviation is 1.5 and 1.6 for key phrases and key sentences, respectively. This indicates that key phrases are as important as key sentences.

- Survey Question 6. *Overall, how would you rate the clustering method in terms of effectively finding distinct main topics of a given Web site? Please give a 1-to-5 score.*

  The overall score of the clustering-summarization framework varies from 3 to 5, leading to an average of 4.1 and a standard deviation of 0.6. This indicates that the summarization framework achieves a reasonable score of 4.1 out of a possible 5.

- Survey Question 7. *What else do you think can be done to improve the summarization system? Any comments, suggestions are more than welcome.*

  Subjects also provide useful comments and feedback from the users' perspective. The following is a list of quotes from the users.

  1. The clustering-summarization approach is much better because it provides more context and is easier to digest. The readers are able to get an idea of what one cluster is about. In contrast, the long summary is too confusing in terms of presenting diverse topics.

2. In the clustering-summarization approach, some of the phrases and sentences seem to be more relevant to the essential topic because they are grouped into one cluster so they seem to make more sense than in a single long summary.
3. It seems that the number of cluster summaries depends on the target Web site and its main topics. For example, the Air Canada Web site can be summarized using less clusters because its scope is smaller than that of the Oracle Web site.
4. It will be better to highlight the main topics if there are more clusters and less elements in each cluster summary.
5. It will be nice to include a heading for each cluster summary so that one knows what that cluster is about.
6. It will be much better to replace personal pronoun (e.g. he) in a key sentence with the real name.
7. It will be better to eliminate service-related pages during the crawling stage. such as copyright pages, which can be long and complicated, but are not very useful for summarizing the main topics.
8. The single long summary takes too much time to comprehend. Thus, it does not seem to be very useful.
9. It is interesting to learn if different clustering methods will lead to different cluster summaries.
10. It is interesting to conduct a study to measure the relatedness of key phrases and key sentences in a cluster summary to topic that the cluster represents.

# Chapter 6

# Conclusion

In this thesis, we propose a framework for summarization of multi-topic Web sites. The system first applies coupled text- and link-based $X$-means clustering to find the most significant topics covered in a given site, and then summarizes each individual cluster using an extraction-based summarization system. Each cluster summary consists of key phrases and key sentences. We conducted a user study to evaluate how well cluster summaries capture the main topics of a given Web site, compared with a single long summary, which is generated using our previous single-topic summarization of an entire Web site. Our user study demonstrates that the clustering-summarization approach statistically significantly outperforms the straightforward summarization approach in the multi-topic Web site summarization task.

The main contribution of this thesis is a framework for clustering and summarization of multi-topic Web sites.

- We demonstrate that the performance of an extraction-based summarization system highly depends on its underlying key phrase extraction method. The C/NC-value key phrase extraction method is demonstrated to be the best out of five alternatives we experimented with.

- We demonstrate that text-based $X$-means clustering with Term Variance feature selection statistically significantly outperforms other clustering configurations in terms of effectively finding the essential topics of a Web site. Moreover, outgoing links can be used to enhance the clustering quality if cross links, as determined by the breadth-first site traversal, are sufficiently rich.

- We propose a classification approach to finding the key sentences in the cluster summarization task. The classifier uses statistical and linguistic features to determine the topical significance of a sentence in addition to the traditional method, where sentences are extracted based on the density of key phrases.

- Intrinsic evaluation is performed on the cluster summaries. Subjects judge how effectively the cluster summaries capture the essential topics of a Web site.

The proposed summarization framework has many potential applications, such as effective organization of search engine results and faceted browsing of large Web sites.

Future research issues include the following:

- Incorporation of more Web-specific features such as availability of phrases in meta data and anchor text to improve key phrase extraction from Web document corpora.
- Investigation of whether anchor text can be helpful in the Web page clustering task.
- Evaluation of clustering using more advanced measures [28].
- Investigation of the subject learning factor to determine whether there is a statistically significant difference within human subjects.
- Evaluation of other factors in multi-document summarization such as coherence, completeness, redundancy, and compression rate.
- Investigation of more advanced learning algorithms such as Support Vector Machines in the key sentence classification task.
- Hierarchical summarization of a Web site to construct a concept hierarchy, which can be obtained by first creating a link hierarchy of the target Web site, and then identifying concept groups of related documents that are further summarized using a coupled content-link summarization system.
- Investigation of applications of our clustering-summarization framework, for example integration with Web site content management systems.
- Experimental comparison of our framework with existing multi-document summarization systems such as MEAD [60].

# Bibliography


[1] J. Allan and H. Raghavan. Using Part-of-speech Patterns to Reduce Query Ambiguity. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, Tampere, Finland, August 11–15, 2002.

[2] E. Amitay and C. Paris. Automatically Summarising Web sites: Is There a Way Around It? In *Proceedings of the Ninth ACM International Conference on Information and Knowledge Management*, pages 173–179, McLean, VA, USA, November 6–11, 2000.

[3] A. Berger and V. Mittal. OCELOT: A System for Summarizing Web Pages. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 144–151, Athens, Greece, July 24–28 2000.

[4] D. Boley. Principal Direction Divisive Partitioning. *Data Mining and Knowledge Discovery*, 2(4):325–344, December 1998.

[5] E. Brill. A Simple Rule-Based Part of Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 152–155, Trento, Italy, March 31–April 3 1992.

[6] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the Seventh International World Wide Web Conference*, pages 107–117, Brisbane, Australia, April 14–18, 1998.

[7] A. Broder, S. Glassman, M. Manasse, and G. Zweig. Syntactic Clustering of the Web. In *Proceedings of the Sixth International World Wide Web Conference*, pages 1157–1166, Santa Clara, CA, USA, April 7–11, 1997.

[8] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices. In *Proceedings of the Tenth International World Wide Web Conference*, pages 652–662, Hong Kong, China, May 01–05, 2001.

[9] S. Chakrabarti, B. Dom, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the Web's Link Structure. *IEEE Computer*, 32(8):60–67, August 1999.

[10] S. Chakrabarti, M. Joshi, and V. Tawde. Enhanced Topic Distillation Using Text, Markup Tags, and Hyperlinks. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 208–216, New Orleans, LA, USA, September 9–12, 2001.

[11] W. Chuang and J. Yang. Extracting Sentence Segments for Text Summarization: A Machine Learning Approach. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 152–159, Athens, Greece, July 24–28, 2000.

[12] D. Cutting, D. Karger, J. Pedersen, and J. Tukey. Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections. In *Proccedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329, Copenhagen, Denmark, June 21–24, 1992.

[13] J. Dean and M. Henzinger. Finding Related Pages in the World Wide Web. In *Proceedings of the Eighth International World Wide Web Conference*, pages 389–401, Toronto, ON, Canada, May 11–14, 1999.

[14] J. Delort, B. Bouchon-Meunier, and M. Rifqi. Enhanced Web Document Summarization Using Hyperlinks. In *Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia*, pages 208–215, Nottingham, UK, August 26–30, 2003.

[15] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38, 1977.

[16] H. Edmundson. New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 16(2):264–285, April 1969.

[17] A. Feelders and W. Verkooijen. Which Method Learns the Most from Data? Methodological Issues in the Analysis of Comparative Studies. In *Proceedings of Fifth International Workshop on Artificial Intelligence and Statistics*, pages 219–225, Ft. Lauderdale, FL, USA, January 4–7, 1995.

[18] G. Flake, S. Lawrence, and L. Giles. Efficient Identification of Web Communities. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, USA, August 20–23, 2000.

[19] C. Fox. Lexical Analysis and Stoplists. In *Information Retrieval: Data Structures and Algorithms*, pages 102–130, 1992.

[20] W. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, June 12, 1992. ISBN: 0-13-463837-9.

[21] E. Frank, G. Paynter, I. Witten, C Gutwin, and C Nevill-Manning. Domain-specific Keyphrase Extraction. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 668–673, Stockholm, Sweden, July 31–August 06, 1999.

[22] K. Frantzi, S. Ananiadou, and H. Mima. Automatic Recognition of Multi-word Terms: the C-value/NC-value Method. *International Journal on Digital Libraries*, 3(2):115–130, August 2000.

[23] E. Glover, K. Tsioutsiouliklis, S. Lawrence, D. Pennock, and G. Flake. Using Web Structure for Classifying and Describing Web Pages. In *Proceedings of the Eleventh International World Wide Web Conference*, pages 562–569, Honolulu, HI, USA, May 07–11, 2002.

[24] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 121–128, Berkeley, CA, USA, August 15–19, 1999.

[25] J. Goldstein, V. Mittal, J. Carbonell, and J. Callan. Creating and Evaluating Multi-document Sentence Extract Summaries. In *Proceedings of the Ninth ACM International Conference on Information and Knowledge Management*, pages 165–172, McLean, VA, USA, November 6–11, 2000.

[26] C. Gutwin, G. Paynter, I. Witten, C. Nevill-Manning, and E. Frank. Improving Browsing in Digital Libraries with Keyphrase Indexes. *Journal of Decision Support Systems*, 27(1–2):81–104, November 1999.

[27] U. Hahn and I. Mani. The Challenges of Automatic Summarization. *IEEE Computer*, 33(11):29–36, November 2000.

[28] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 17(2–3):107–145, December 2001.

[29] V. Hatzivassiloglou, J. Klavans, M. Holcombe, R. Barzilay, M. Kan, and K. McKeown. Simfinder: A Flexible Clustering Tool for Summarization. In *Proceedings of the NAACL'01 Workshop on Automatic Summarization*, pages 41–49, Pittsburgh, PA, USA, June 3, 2001.

[30] A. Hotho, A. Maedche, and S. Staab. Ontology-Based Text Clustering. In *Proceedings of the IJCAI'01 Workshop on "Text Learning: Beyond Supervision"*, Seattle, WA, USA, August 6, 2001.

[31] S. Jones, S. Lundy, and G. Paynter. Interactive Document Summarisation Using Automatically Extracted Keyphrases. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences, Volume 4*, Big Island, Hawaii, January 7–10, 2002.

[32] S. Jones and G. Paynter. Automatic Extraction of Document Keyprases for Use in Digital Libraries: Evaluation and Applications. *Journal of the American Society for Information Science and Technology*, 53(8):653–677, April 2002.

[33] S. Jones and G. Paynter. An Evaluation of Document Keyphrase Sets. *Journal of Digital Information*, 4(1), February 19, 2003.

[34] M. Kessler. Bibliographic Coupling between Scientific Papers. *American Documentation*, 14:10–25, 1963.

[35] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, Septermber 1999.

[36] J. Kogan, C. Nicholas, and V. Volkovich. Text Mining with Information-Theoretic Clustering. *Computing in Science and Engineering*, 5(6):52–59, November/December 2003.

[37] B. Krulwich and C. Burkey. Learning User Information Interests through the Extraction of Semantically Significant Phrases. In *AAAI Spring Symposium Technical Report SS-96-05: Machine Learning in Information Access*, pages 110–112, 1996.

[38] J. Kupiec, J. Pedersen, and F. Chen. A Trainable Document Summarizer. In *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73, Seattle, WA, USA, July 09–13, 1995.

[39] D. Lawrie, B. Croft, and A. Rosenberg. Finding Topic Words for Hierarchical Summarization. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 349–357, New Orleans, LA, USA, September 9–12, 2001.

[40] D. Lawrie and W. Croft. Generating Hierarchical Summaries for Web Searches. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 457–458, Toronto, ON, Canada, July 28–August 1, 2003.

[41] W. Li, N. Ayan, O. Kolak, Q. Vu, H. Takano, and H. Shimamura. Constructing Multi-Granular and Topic-Focused Web Site Maps. In *Proceedings of the Tenth International World Wide Web Conference*, pages 343–354, Hong Kong, China, May 1–5, 2001.

[42] C. Lin and E. Hovy. From Single to Multi-document Summarization: A Prototype System and its Evaluation. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, pages 457–464, Philadelphia, PA, USA, July 7–12, 2002.

[43] B. Liu and K. Chang. Editorial: Special Issue on Web Content Mining. *ACM SIGKDD Explorations Newsletter*, 6(2):1–4, December 2004.

[44] H. Liu and L. Yu. Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502, April 2005.

[45] T. Liu, S. Liu, Z. Chen, and W. Ma. An Evaluation on Feature Selection for Text Clustering. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 488–495, Washington, DC, USA, August 21-24, 2003.

[46] W. Lu, J. Janssen, E. Milios, N. Japkowicz, and Y. Zhang. Node Similarity in the Citation Graph. *Knowledge and Information Systems: An International Journal*, 11(1):105–129, January 2007. Available at `http://dx.doi.org/10.1007/s10115-006-0023-9`, last visited on February 12, 2007.

[47] H. Luhn. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159–165, April 1958.

[48] I. Mani. Recent Developments in Text Summarization. In *Proceedings of the Tenth ACM International Conference on Information and Knowledge Management*, pages 529–531, Atlanta, GA, USA, November 5–10, 2001.

[49] I. Mani, D. House, G. Klein, L. Hirschman, T. Firmin, and B. Sundheim. The TIPSTER SUMMAC Text Summarization Evaluation. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 77–85, Bergen, Norway, June 8–12, 1999.

[50] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, June 18 1999. ISBN: 0-262-13360-1.

[51] D. Marcu. From Discourse Structures to Text Summaries. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 82–88, Madrid, Spain, July 1997.

[52] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, M. Kan, B. Schiffman, and S. Teufel. Columbia Multi-document Summarization: Approach and Evaluation. In *Proceedings of the Workshop on Text Summarization of the Document Understanding Conference*, New Orleans, LA, USA, September 13–14, 2001.

[53] K. McKeown, R. Passonneau, D. Elson, A. Nenkova, and J. Hirschberg. Do Summaries Help? A Task-Based Evaluation of Multi-Document Summarization. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 210–217, Salvador, Brazil, August 15–19, 2001.

[54] E. Milios, Y. Zhang, B. He, and L. Dong. Automatic Term Extraction and Document Similarity in Special Text Corpora. In *Proceedings of the Sixth Conference of the Pacific Association for Computational Linguistics*, pages 275–284, Halifax, NS, Canada, August 22–25, 2003.

[55] T. Mitchell. *Machine Learning.* Mcgraw-Hill International Edit, 1997.

[56] D. Modha and W. Spangler. Clustering Hypertext with Applications to Web Searching. In *Proceedings of the Eleventh ACM Conference on Hypertext and Hypermedia*, pages 143–152, San Antonio, TX, USA, May 30–June 3, 2000.

[57] H. Nakagawa. Experimental Evaluation of Ranking and Selection Methods in Term Extraction. In D. Bourigault, C. Jacquemin, and M. L'Homme, editors, *Recent Advances in Computational Terminology*, pages 303–325, 2001.

[58] D. Pelleg and A. Moore. $X$-means: Extending $K$-means with Efficient Estimation of the Number of Clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, Standord, CA, USA, June 29–July 2, 2000.

[59] D. Radev, W. Fan, H. Qi, H. Wu, and A. Grewal. Probabilistic Question Answering on the Web. In *Proceedings of the Eleventh International World Wide Web Conference*, pages 408–419, Honolulu, Hawaii, USA, May 7–11, 2002.

[60] D. Radev, H. Jing, and M. Budzikowska. Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies. In *Proceedings of the ANLP'00/NAACL'00 Workshop on Automatic Summarization*, pages 21–29, Seattle, WA, USA, April 2000.

[61] G. Salton and C. Buckley. Term Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[62] G. Salton and M. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc., New York, NY, USA, 1986. ISBN: 0070544840.

[63] M. Sanderson and B. Croft. Deriving Concept Hierarchies from Text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–213, Berkeley, CA, USA, August 15–19, 1999.

[64] J. Schlesinger, J. Conroy, M. Okurowski, and D. O'Leary. Machine and Human Performance for Single and Multidocument Summarization. *IEEE Intelligent Systems*, 18(1):46–54, January/February 2003.

[65] H. Schütze and H. Silverstein. Projections for Efficient Document Clustering. In *Proceedings of the Twentieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–81, Philadelphia, PA, USA, July 27–31, 1997.

[66] F. Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1–47, March 2002.

[67] M. Sinka and D. Corne. Towards Modernized and Web-Specific Stoplists for Web Document Analysis. In *Proccedings of the IEEE/WIC International Conference on Web Intelligence*, pages 396–402, Halifax, NS, Canada, October 13–16, 2003.

[68] H. Small. Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents. *Journal of the American Society for Information Science*, 24(4):265–269, 1973.

[69] K. Sparck-Jones and J. Galliers. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer-Verlag New York, Inc., New York, NY, USA, June 1996. ISBN 3-540-61309-9.

[70] E. Spertus. ParaSite: Mining Structural Information on the Web. In *Proceedings of the Sixth International World Wide Web Conference*, pages 1205–1215, Santa Clara, CA, USA, April 7–11, 1997.

[71] G. Stein, A. Bagga, and G. Wise. Multi-document Summarization: Methodologies and Evaluations. In *Proceedings of the Seventh Conference on Automatic Natural Language Processing*, pages 337–346, Lausanne, Switzerland, October 2000.

[72] M. Steinbach, G. Karypis, and V. Kumar. A Comparison of Document Clustering Techniques. In *Proceedings of SIGKDD'00 Workshop on Text Mining*, Boston, MA, USA, August 20, 2000.

[73] S. Teufel and M. Moens. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, 28(4):409–445, 2002.

[74] P. Turney. Extraction of Keyphrases from Text: Evaluation of Four Algorithms. Technical Report ERB-1051 (NRC-41550), Institute for Information Technology, National Research Council of Canada, Ottawa, ON, Canada, October 23, 1997. Available at `http://citeseer.ist.psu.edu/turney97extraction.html`, last visited on November 26, 2004.

[75] P. Turney. Learning Algorithms for Keyphrase Extraction. *Information Retrieval*, 2(4):303–336, May 2000.

[76] P. Turney. Mining the Web for Lexical Knowledge to Improve Keyphrase Extraction: Learning from Labeled and Unlabeled Data. Technical Report ERB-1096 (NRC-44947), Institute for Information Technology, National Research Council of Canada, Ottawa, ON, Canada, August 13, 2002. Available at `http://citeseer.ist.psu.edu/turney02mining.html`, last visited on November 26, 2004.

[77] P. Turney. Coherent Keyphrase Extraction via Web Mining. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 434–439, Acapulco, Mexico, August 9–15, 2003.

[78] Y. Wang and M. Kitsuregawa. Use Link-Based Clustering to Improve Web Search Results. In *Proceedings of the Second International Conference on Web Information Systems Engineering*, pages 115–124, Kyoto, Japan, December 3–6, 2001.

[79] Y. Wang and M. Kitsuregawa. Evaluating Contents-link Coupled Web Page Clustering for Web Search Results. In *Proceedings of the Eleventh ACM International Conference on Information and Knowledge Management*, pages 499–506, McLean, VA, USA, November 04–09, 2002.

[80] Wikipedia. Inter-rater Reliability. Available at `http://en.wikipedia.org/wiki/Inter-rater_reliability`, last visited on December 6, 2006.

[81] I. Witten, G. Paynter, E. Frank, C. Gutwin, and C. Nevill-Manning. KEA: Practical Automatic Keyphrase Extraction. In *Proceedings of the Fourth ACM Conference on Digital Libraries*, pages 254–255, Berkeley, CA, USA, August 11–14, 1999.

[82] Y. Yang and J. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420, Nashville, TN, USA, July 8–12, 1997.

[83] Y. Zhang, E. Milios, and N. Zincir-Heywood. A Comparison of Keyword- and Keyterm-Based Methods for Automatic Web Site Summarization. Technical Report CS-2004-11, Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada, October 2, 2004. Available at `http://www.cs.dal.ca/research/techreports/2004/CS-2004-11.shtml`, last visited on November 26, 2004.

[84] Y. Zhang, N. Zincir-Heywood, and E. Milios. Term-Based Clustering and Summarization of Web Page Collections. In *Advances in Artificial Intelligence, Proceedings of the Seventeenth Conference of the Canadian Society for Computational Studies of Intelligence*, pages 60–74, London, ON, Canada, May 17–19, 2004.

[85] Y. Zhang, N. Zincir-Heywood, and E. Milios. World Wide Web Site Summarization. *Web Intelligence and Agent Systems: An International Journal*, 2(1):39–53, June 2004.

[86] Y. Zhang, N. Zincir-Heywood, and E. Milios. Narrative Text Classification for Automatic Key Phrase Extraction in Web Document Corpora. In *Proceedings of the Seventh ACM International Workshop on Web Information and Data Management*, pages 51–58, Bremen, Germany, November 5, 2005.

[87] Y. Zhao and G. Karypis. Evaluation of Hierarchical Clustering Algorithms for Document Datasets. In *Proceedings of the Eleventh ACM International Conference on Information and Knowledge Management*, pages 515–524, McLean, VA, USA, November 04–09, 2002.

# Appendix A

# Web Page Parsing

In this chapter, we discuss how to process a Web page URL when the page is parsed. During the parsing process, the following tasks will be performed: link filtering, link resolution, link validation, link pruning, and file truncation.

## A.1 Link Filtering

We aim to focus on static Web pages only and filter out noisy URLs. Link filtering does the following:

- Remove all query terms in a URL because we focus on static pages, not dynamic pages. For instance, `http://www.dal.ca/search.html?Computer+Science` will be truncated into `http://www.dal.ca/search.html`
- Remove the fragment (also known as reference) part, which is indicated by the sharp sign character "#", because a fragment only indicates that after the specified resource is retrieved, the application is specifically interested in that part of the document. For example, `http://www.cs.dal.ca/grad.html#Resources` will be truncated into `http://www.cs.dal.ca/grad.html`.
- Discard a URL if it contains one or more of the following punctuation marks, i.e., ' ! @ $ ∧ & * ( ) = [ ] { } — \ ; ' < > , and blank space.

## A.2 Link Resolution

Convert a relative URL against the base URL into an absolute URL. For example, `/course` against the FCS home page `http://www.cs.dal.ca/` yields `http://www.cs.dal.ca/course`.

## A.3 Link Validation

We aim to focus on Web pages of the **text/html** content type from the same host. Link validation does the following:

- A URL is tested whether it has the same host as the homepage. We only look at in-host Web pages and investigation of out-host pages is a future research direction.

- Test whether there are connection and access problems such as the "Page not found 404" error.

- For URLs with new port numbers, we treat them as external URLs from other hosts and discard them accordingly. For example, when traversing `http://www.cs.dal.ca` (default port: 80), we find a URL `http://www.cs.dal.ca:6000`, which is treated as a out-host page. Consequently this page will not be further visited and parsed.

- For URLs of usernames such as `http://www.cs.dal.ca/~user`, they are not added into Q (queue of pages to visit), which means that they will not be further visited and parsed. However, these pages are still part of the set of outgoing links of the target page and they are treated as leaves in the final link hierarchy.

- In order to avoid parsing either a known binary file (for example, .jpg, .bmp, .doc and .dat), or a file whose file type is non-obvious or unfamiliar to us, we only deal with Web pages with the **text/html** content type. All Web pages of other file types will be treated as binary files. They are added into the link hierarchy but not Q, the queue of pages to visit. Again, these pages are only part of outgoing links of a particular page and treated as leaf nodes in the link hierarchy.

  Conversion of .pdf files into pure text for text analysis is a future research direction because .pdf files are mostly documentation records or forms and therefore they are too specific for summary generation. Omitting them should not have a big effect on quality of summarization.

## A.4 Link Pruning

For each page, the set of its outgoing links will be examined in order to select only a proper subset of links. For example, when parsing the page `http://www.cs.dal.ca/news`, we find many links of specific news such as `http://www.cs.dal.ca/news/2005-03-01.shtml`. These pages very likely are leaves in the link hierarchy. If the number of such links is too big, say 500, then most probably inclusion of all these pages does not benefit the summarization. Thus, we select only the first 10 representatives of such pages. Of course, links with further paths such as `http://www.cs.dal.ca/news/archive/` are added to the link hierarchy.

## A.5 File Truncation

For each URL found in a page, we check the size of the page represented by this URL. If it is greater than 50K, we truncate it to that size. The reason is that some documents such as privacy statements contain too much text information and use of their full text does not benefit summarization at all.

# Appendix B

# User Study Instructions

This chapter presents the instructions on how to conduct the user study. It includes a summary rating example and a short survey.

## B.1 A Summary Rating Example

Each user is provided with a summary rating example, which is designed to help them be familiar with the evaluation process. A snippet of the summary rating example is shown in Figure B.1.

## B.2 A Short Survey

At the end of the user study, users have to complete a short survey, which is designed to get feedback on improving our summarization systems. There are seven survey questions as follows:

1. Overall, which type of summary do you like better, the single long summary or the top five short cluster summaries?

2. If the answer to Question 1 is the latter, how many cluster summaries do you prefer?

3. For the single long summary, what is the ideal number of key phrases in your opinion? And key sentences?

4. For each cluster summary, do the same as in Question 3.

5. The summary consists of two parts, key phrases and key sentences. Which part do you think is more important? Suppose there are 10 points of importance for the whole summary. How many points should be assigned to each part (e.g. 4 to key phrases and 6 to key sentences), respectively?

A Summary Rating Example for SEI: http://www.sei.cmu.edu

Ranking scales
5: strongly related; 4: well related; 3: fairly related; 2: poorly related; 1: not related.

The most essential topic extracted
Software Engineering Institute at CMU for improvement of software engineering management and practice.

| Top 25 key phrases | Ratings |
|---|---|
| 1) software engineering institute | 5 |
| 2) engineering institute | 2 |
| 3) software engineering | 5 |
| 4) carnegie mellon university | 4 |
| 5) carnegie mellon | 2 |
| ..... | |

| Top 25 key sentences | Ratings |
|---|---|
| 1) The Software Engineering Institute (SEI) sponsors, co-sponsors, and is otherwise involved in many events throughout the year. | 4 |
| 2) The Software Engineering Institute offers a number of courses and training opportunities. | 4 |
| 3) Information contained on the Software Engineering Institute Web site is published in the interest of scientific and technical information exchange. | 3 |
| 4) The Software Engineering Institute is a federally funded research and development center sponsored by the U.S. Department of Defense and operated by Carnegie Mellon University. | 5 |
| 5) As CEO and director, Nielsen's responsibilities will include setting a technical and business strategy for the Software Engineering Institute. | 3 |
| ..... | |

Figure B.1: Snippet of a summary rating example.

6. Overall, how would you rate the clustering method in terms of effectively finding distinct main topics of a given Web site? Please give a 1-to-5 score.

7. What else do you think can be done to improve the summarization system? Any comments, suggestions are more than welcome.

# Appendix C

## A Full List of Summaries

This chapter presents both the single long summary and the top five short cluster summaries for each of the six test Web sites.

1. `http://www.sei.cmu.edu`

- **A Single Long Summary**

  Part I. Top 25 Key Phrases

  - software engineering institute
  - engineering institute
  - software engineering
  - carnegie mellon university
  - carnegie mellon
  - mellon university
  - carnegie mellon university term
  - capability maturity model
  - capability maturity model integration
  - capability maturity
  - maturity model
  - maturity model integration
  - software process
  - software engineering institute partner
  - model integration
  - product line

- engineering institute partner
- team software process
- institute partner
- software architecture
- personal software process
- general navigation button
- process improvement
- topic navigation button
- team software

Part II. Top 25 Key Sentences

- The Software Engineering Institute (SEI) sponsors, co-sponsors, and is otherwise involved in many events throughout the year.
- The Software Engineering Institute offers a number of courses and training opportunities.
- Information contained on the Software Engineering Institute Web site is published in the interest of scientific and technical information exchange.
- The Software Engineering Institute is a federally funded research and development center sponsored by the U.S. Department of Defense and operated by Carnegie Mellon University.
- As CEO and director, Nielsen's responsibilities will include setting a technical and business strategy for the Software Engineering Institute.
- The Software Engineering Institute recently provided the Internal Revenue Service with an independent report on the IRS's delayed, over-cost modernization effort.
- Nielsen will take over as CEO and director of the Software Engineering Institute on August 1.
- IEEE Spectrum: Why Software Fails September 2005 Article about the Capability Maturity Model (CMM) created by the Software Engineering

Institute to help organizations assess and analyze their software development practices.

- Nielsen is selected as CEO and director of the Software Engineering Institute.
- Len Bass is a Senior Member of the Technical Staff at the Software Engineering Institute (SEI) and participates in the High Dependability Computing Program.
- Peter is a senior member of the technical staff member at the Software Engineering Institute (SEI).
- Jorgen Hansson is a senior member of the technical staff at the Software Engineering Institute (SEI).
- Thomas Longstaff is the Deputy Director for Technology in the Networked Systems Survivability (NSS) Program at the Software Engineering Institute (SEI).
- Prior to coming to the Software Engineering Institute, Longstaff was the technical director at the Computer Incident Advisory Capability (CIAC) at Lawrence Livermore National Laboratory in Livermore, California.
- Mead is a senior member of the technical staff in the Networked Systems Survivability Program at the Software Engineering Institute (SEI).
- Shimeall is a Senior Member of the Technical Staff with the Networked Systems Survivability Program at the Software Engineering Institute (SEI).
- At the Software Engineering Institute (SEI), we have been working in open systems since 1993, developing courses, related products, and other sources of open systems information.
- news@sei is a quarterly hardcopy newsletter published by the Software Engineering Institute that gives readers an overview of SEI work, events and publications.
- Paul Clements is a Senior Member of the Technical Staff, Software Engineering Institute.

- The Product Line Systems Program of the Software Engineering Institute is proud to be represented by the following books.
- The Software Engineering Institute has established a software architecture curriculum.
- A 1996 technical report [1] from the Software Engineering Institute, in its introduction points out that software architecture is still a relatively immature area from both research and practice perspectives.
- Video lectures for this course include resident experts at the Software Engineering Institute and recognized leaders from industry and academia.
- Humphrey describes a software process improvement strategy based on the software process maturity model developed at the Software Engineering Institute.
- This report presents the interim results of work done by members of the Networked Systems Survivability Program at the Software Engineering Institute in exploring these issues.

- **Top 5 Short Cluster Summaries**

  Cluster I. Software Architecture

  - software architecture
  - software engineering
  - software architecture professional
  - architecture professional
  - professional certificate
  - SEI architecture experts provide technical assistance and coaching in software architecture requirements, software architecture design, software architecture documentation, and architecture-centric life-cycle practices.
  - He is a co-author of Applied Software Architecture and Documenting Software Architectures: Views and Beyond, published by Addison-Wesley and lectures on architecture-centric approaches.

- The significance of the program family concept to software architecture is that software architecture embodies those decisions at or near the top of Parnas' program family decision tree.
- Raghuraman Ramasubbu(Senior Software Engineer, Mastech Corporation): Software architecture is a framework that provides the basis for manifestation of all software objects within an enterprise.
- Human Aspects of Software Engineering details software engineering from the perspective of those involved in the software development process: individuals, team, customers, and the organization.

Cluster II. Open System

- open system
- process research
- process improvement
- software process
- software engineering research
- This tutorial covers key open system concepts and definitions, an open system engineering process, open system policy, conformance management, todays transition environment, and an open system transition process.
- The IPRC is not trying to solve a particular business problem, but to chart potential directions for the future of software process research.
- We are sponsoring the IPRC as a focal point for top researchers and forward-thinking organizations investigating the latest in software process research.
- We are part of a world recognized center of excellence for software engineering research, the Software Engineering Institute, which serves as a trusted broker among industry, government, and academia.
- If the Engineering Process group is new to process improvement, members should consider taking the Defining Software Processes or Mastering Process Improvement courses.

Cluster III. Risk Management

- risk management
- software engineering
- continuous risk
- continuous risk management
- software engineering institute
- Customers/collaborators are welcomed for work that will further refine the areas of Software Risk Evaluation, Continuous Risk Management (CRM), Team Risk Management, and Risk Process Checks.
- The Team Risk Management Service extends Continuous Risk Management overview to all organizations in a program, tailoring methods and tools to the joint management of program risks.
- The Continuous Risk Management Guidebook was developed to help a project or organization establish continuous risk management as a routine practice and then continue to improve this process.
- The SEI's products for Continuous Risk Management are completely consistent with the requirements of RSKM.
- Successful team risk management depends upon having implemented, or partially implemented, Continuous Risk Management within the program's organizations.

Cluster IV. Product Line

- product line
- software product line
- software product
- line practice
- product line practice
- Initiating and operating a software product line organization is a low-risk, predictable process that results in little or no organizational upheaval.

- Third-party vendors provide tools and services that support software product lines.
- Influential industry analysts promote software product lines as a critical technology.
- Government acquisition mangers are well educated about software product lines.
- The organizational interest in software product line has grown tremendously over the last five years.

Cluster V. People CMM

- people cmm
- software engineering institute
- capability maturity model
- maturity model
- capability maturity
- This course is a prerequisite for the Intermediate Concepts of the People CMM course, People CMM Assessment Team Training, and People CMM Lead Appraiser Training.
- The People Capability Maturity Model (People CMM) is a framework that helps organizations successfully address their critical people issues.
- Gians work at the SEI includes assisting organizations in successfully addressing their critical human capital issues through the use of the SEIs People Capability Maturity Model (People CMM).
- The People Capability Maturity Model (People CMM) is a framework that guides organizations in improving their processes for managing and developing their workforces.
- Systems integrators worldwide are adopting the Software Engineering Institute's Capability Maturity Model Integration, a technology that is replacing the traditional Capability Maturity Model.

2. `http://www.aiai.ed.ac.uk`

- **A Single Long Summary**

  Part I. Top 25 Key Phrases

  - artificial intelligence applications institute
  - artificial intelligence application
  - version date source
  - artificial intelligence
  - intelligence application
  - version date
  - plan representation
  - knowledge acquisition
  - process interchange format
  - knowledge based system
  - activity representation
  - ai planning
  - core group
  - process specification language
  - nist process specification language
  - knowledge management
  - planning process
  - core plan representation
  - case based reasoning
  - spar core group
  - specification language
  - knowledge representation
  - process interchange

- air force
- activity specification

Part II. Top 25 Key Sentences

- Back to Home page Artificial Intelligence Applications Institute School of Informatics, The University of Edinburgh.
- IM-PACs is be based on I-X, a key research orientated asset of the Artificial Intelligence Applications Institute (AIAI) at the University of Edinburgh.
- Dr Stuart Aitken Artificial Intelligence Applications Institute The University of Edinburgh Edinburgh EH8 9LE Email: stuart@aiai.ed.ac.uk
- WxCLIPS was started in 1992 by Julian Smart at the Artificial Intelligence Applications Institute, part of the University of Edinburgh.
- I-X is a valuable asset of the Artificial Intelligence Applications Institute and must not be used without the prior permission of a rights holder.
- O-Plan is a valuable asset of the Artificial Intelligence Applications Institute and must not be used without the prior permission of a rights holder.
- Professor Austin Tate, Artificial Intelligence Applications Institute, Division of Informatics, University of Edinburgh.
- Artificial Intelligence Applications Institute, University of Edinburgh.
- Process Steps, Process Products and System Capabilities, Artificial Intelligence Applications Institute, University of Edinburgh, Technical Report ISAT-AIAI/TR/4 Version 2, April 14, 1997.
- Rob was a pioneer of artificial intelligence applications, creating a successful company based in Livingston.
- Rob Milne was perhaps the best-known and most successful person in Europe who has produced and promoted artificial intelligence applications.
- Analysis of Candidate PSL Process/Plan Representations, Artificial Intelligence Applications Institute, AIAI-PR-66, Edinburgh, Scotland.

- EXPECT: Explicit Representations for Flexible Acquisition In Proceedings of the Ninth Knowledge Acquisition for Knowledge-Based Systems Workshop, February 26-March 3, 1995.
- For example, SPAR has been a contributing source towards the development of the Core Plan Representation (CPR) [Pease & Carrico, 1997].
- As an example of the level at which the Reference Object Model for SPAR will be described, the OMWG Core Plan Representation [Pease & Carrico, 1997] Object Model is shown here.
- As an example of the level at which the Reference Object Model will be provided, see the current OMWG Core Plan Representation Request For Comment document [Pease & Carrico, 1997].
- The working group members represent some of the most experienced people worldwide who have been involved in creating shared plan, process and activity representations or standards for some years.
- The paper provides a comprehensive bibliography and related world wide web resources for work in the area of plan, process and activity representation.
- AIAI is a collaborator on the National Institute for Standards and Technology (NIST) Process Specification Language (PSL) project.
- The work on SPAR has been conducted alongside the development of the National Institute of Standards and Technology Process Specification Language (NIST PSL).
- Advanced Knowledge Technologies Rapid Knowledge Formation Case-based Reasoning Case-based Reasoning has been successfully applied to fraud detection in the finance industry.

- **Top 5 Short Cluster Summaries**

  Cluster I. Intelligence Application

  - intelligence application
  - artificial intelligence applications institute

- intelligence applications institute
- artificial intelligence application
- applications institute
- IM-PACs is be based on I-X, a key research orientated asset of the Artificial Intelligence Applications Institute (AIAI) at the University of Edinburgh.
- Rob was a pioneer of artificial intelligence applications, creating a successful company based in Livingston.
- WxCLIPS was started in 1992 by Julian Smart at the Artificial Intelligence Applications Institute, part of the University of Edinburgh.
- I-X is a valuable asset of the Artificial Intelligence Applications Institute and must not be used without the prior permission of a rights holder.
- O-Plan is a valuable asset of the Artificial Intelligence Applications Institute and must not be used without the prior permission of a rights holder.

Cluster II. Knowledge Based System

- knowledge based system
- knowledge base
- knowledge acquisition
- knowledge engineering
- knowledge modelling
- Areas of expertise include the use of methodologies for knowledge engineering, knowledge acquisition techniques, and in programming knowledge-based software systems.
- The fundamental training strategy behind this "journeyman" scheme is the building of a fully documented knowledge based system with supervision from AIAI staff.
- Increasingly, belief network techniques are being employed to deliver advanced knowledge based systems to solve real world problems.

- The approach was tested and refined through the development of a knowledge-based system for the support of planning for search and rescue.
- Edinburgh have developed the Enterprise ontology for representing an organization and its activities, as well as having considerable experience in applying knowledge modelling to the knowledge engineering process.

Cluster III. Intelligent System

- intelligent system
- case based reasoning
- intelligent systems creation
- system development
- system development process
- (2006) The Helpful Environment, Special Issue on Next 50 Years of Intelligent Systems, IEEE Intelligent Systems.
- Advanced Knowledge Technologies Rapid Knowledge Formation Case-based Reasoning Case-based Reasoning has been successfully applied to fraud detection in the finance industry.
- AIAI can play a role in Intelligent System development projects during any of these phases, but most typically would be involved during the earlier stages of analysis, modelling and design, and proof-of-concept implementation where AIAI's experience of working with specifically AI concepts and technologies will be most telling.
- Systems Integration - A broad vision of an open architecture for intelligent systems creation for synthesis tasks (such as planning, design and configuration) based on the handling of "issues" and the management or maintenance of the constraints describing the product of the process.
- During the first year of AUSDA, AIAI will develop a generic framework for modelling the system development process.

Cluster IV.

- ai planning
- planning system
- ai planning system
- plan representation
- plan ontology
- Responsive Planning and Scheduling Using AI Planning Techniques - Optimum-AIV - Austin Tate - in "Trends & Controversies - AI Planning Systems in the Real World", IEEE Expert: Intelligent Systems & their Applications, Vol. 11 No. 6, pp. 4-12, December 1996
- Dr John Levine is an Informatics Research Fellow working on evolutionary computation and AI planning systems.
- The Nonlin hierarchical partial-order AI planning system, developed by Austin Tate at the University of Edinburgh, is available in a browsable and downloadable form here:

Cluster V.

- air force
- air force research
- air force research laboratory
- force research laboratory
- research laboratory
- This work is sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (Rome), Air Force Materiel Command, USAF, under grant numbers F30602-95-1-0022 and F30602-97-C-0275.
- AIAI is part of the DARPA/Air Force Research Laboratory (Rome) Planning Initiative (ARPI) Initiative Support and ACPT Testbed (ISAT) Project, working as a subcontractor to ISX Corporation.

- Supported by The Technical Cooperation Program, United States Air Force Research Laboratory/Information Directorate, Rome Research Site, Defense Advanced Research Projects Agency, UK Defence Science and Technology Laboratory, and Defence Science and Technology Organsation, Australia.
- From 1998 to 2000, the O-Plan project was supported by the the US Planning Initiative/Planning and Decision Aids Program (ARPI/PDA), the US Army Small Unit Operations Program (SUO), and the Cooperating Agent Based Systems Program (CoABS) - with funding through the Defense Advanced Resaerch Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) by contract number F30602-99-1-0024 monitored by Wayne Bosco at AFRL.
- In August 1997, DARPA and the Air Force Research Laboratory (Rome) Planning Initiative (ARPI) Program Managers proposed an effort to build on the accumulated expertise from past DARPA and ARPI sponsored research in order to create a shared plan representation suitable for use in ARPI and on applied research programmes in their communities.

3. `http://www.ai.uga.edu`

- **A Single Long Summary**

  Part I. Top 25 Key Phrases

  - artificial intelligence
  - artificial intelligence center
  - ai center
  - university system
  - intelligence center
  - natural language
  - programming language
  - baud rate

- natural language processing
- assembly language
- c compiler
- program memory
- neural network
- operating system
- degree program
- development tool
- external memory
- embedded processor
- embedded controller
- model snake
- independent study
- von neuman architecture
- serial port
- sample code
- sparse ii

Part II. Top 25 Key Sentences

- The Artificial Intelligence Center is an interdepartmental research and instructional center within the Franklin College of Arts and Sciences of the University of Georgia.
- It was officially established as the Artificial Intelligence Center in 1995.
- Located in the same building as the Artificial Intelligence Center, the library has millions of books and bound periodicals and also provides online access to many indexes and journals.
- The Master of Science in Artificial Intelligence (M.S.A.I.) degree program is offered by the interdisciplinary Artificial Intelligence Center.

- Artificial Intelligence Center The University of Georgia 111 Boyd GSRC Athens, Georgia 30606-7415 U.S.A.
- The workshop, hosted by the Artificial Intelligence Center of the University of Georgia will be held April 16, 2004 at the Georgia Center for Continuing Education.
- To see if the author of a package is still at The University of Georgia, use the University's online directory, and if unsuccessful, contact the Artificial Intelligence Center.
- Covington % Artificial Intelligence Center % The University of Georgia % Athens, Georgia 30602-7415 % July 28, 1994
- Opened in June 1999, this is a small electronics laboratory established within the Artificial Intelligence Center to enable faculty and students to participate in research programs involving microelectronics.
- Although the Artificial Intelligence Center does not teach electrical engineering, many students arrive with an electrical engineering background, and this laboratory helps them put it to good use.
- The Artificial Intelligence Center has an Industrial Partners Program giving corporations access to our research and our graduates entering the job market.
- The Artificial Intelligence Center houses two degree programs, the Master of Science program in Artificial Intelligence and the bachelor's degree program in Cognitive Science.
- The AI Center currently offers a two-year interdisciplinary masters degree program in Artificial Intelligence and houses the undergraduate Cognitive Science program for UGA.
- The following is a list of the various anonymous ftp sites that have 8051 source code and programming languages.
- The following is a list of the anonymous ftp sites that have source code and programming languages for various microcontrollers.

- The masters program includes a comprehensive study of AI, with strengths on logic programming, expert systems, neural networks, genetic algorithms, and natural language processing.
- Areas of specialization include automated reasoning, cognitive modeling, neural networks, genetic algorithms, expert databases, expert systems, knowledge representation, logic programming, and natural-language processing.
- Natural Language Processing for Prolog Programmers, by Covington, is being "printed on demand" by Prentice Hall.
- You must do library research to determine how your work fits into the existing science and technology of natural language processing.
- It provides a true introduction to Natural Language Processing without requiring familiarity with Lisp or Prolog.
- Applied Natural Language Processing NOTE: Future plans for LING 6570 are unclear.
- This compiler recognizes a subset of the BASIC programming language and generates assembly language sources files.
- Has a unique architecture with three memory spaces: program memory, data memory, and a CPU register file.
- Program: 8051.zip Description: Many development tools including: debugger, monitor, LCD and stepper moter driver, communications, host client, and much more.

- **Top 5 Short Cluster Summaries**

  Cluster I.

  - computer science
  - artificial intelligence
  - ai center
  - artificial intelligence center
  - intelligence center

- Areas of specialization include automated reasoning, cognitive modeling, neural networks, genetic algorithms, expert databases, expert systems, knowledge representation, logic programming, and natural-language processing.
- The Artificial Intelligence Center is an interdepartmental research and instructional center within the Franklin College of Arts and Sciences of the University of Georgia.
- The Artificial Intelligence Center houses two degree programs, the Master of Science program in Artificial Intelligence and the bachelor's degree program in Cognitive Science.
- It was officially established as the Artificial Intelligence Center in 1995.
- Located in the same building as the Artificial Intelligence Center, the library has millions of books and bound periodicals and also provides online access to many indexes and journals.

Cluster II.

- neural network
- fuzzy logic
- fuzzy logic software
- genetic algorithm
- neural network technology
- Fuzzy Logic and neural networks are two design methods that are coming into favor in embedded systems.
- The masters program includes a comprehensive study of AI, with strengths on logic programming, expert systems, neural networks, genetic algorithms, and natural language processing.
- Areas of specialization include automated reasoning, cognitive modeling, neural networks, genetic algorithms, expert databases, expert systems, knowledge representation, logic programming, and natural-language processing.

- This is the step where neural networks technology can be helpful to the fuzzy-logic designer.
- In an effort to change fuzzy logic from a "buzzword" (as it is in most parts of the world) to a well established design method (as it is in Japan), most manufacturers of microcontrollers have introduced fuzzy logic software.

Cluster III.

- natural language
- language processing
- natural language processing
- artificial intelligence
- programming language
- The masters program includes a comprehensive study of AI, with strengths on logic programming, expert systems, neural networks, genetic algorithms, and natural language processing.
- Areas of specialization include automated reasoning, cognitive modeling, neural networks, genetic algorithms, expert databases, expert systems, knowledge representation, logic programming, and natural-language processing.
- Natural Language Processing for Prolog Programmers, by Covington, is being "printed on demand" by Prentice Hall.
- You must do library research to determine how your work fits into the existing science and technology of natural language processing.
- It provides a true introduction to Natural Language Processing without requiring familiarity with Lisp or Prolog.

Cluster IV.

- embedded system
- embedded systems programming
- embedded system software

- embedded processor
- embedded controller
- Embedded Systems Programming - programming and systems design articles - Miller Freeman Publications - 500 Howard St., San Francisco, CA 94105 - (415) 397-1881
- Embedded Systems Programming in C and Assembler - John Forrest Brown - Van Nostrand Reinhold, 1994 - 304 pages, $49.95 - ISBN 0-442-01817-7 - covers Motorola and Intel processors - includes diskette with code from the book - book review in Dr.
- Based on the article "A Portable Real Time Kernel in C" in Embedded Systems Programming (Part 1: vol 5 no 5 May 1992, Part 2: vol 5 no 6 June 1992)
- Embedded Systems Programming Periodical, focus on embedded system software issues.
- Is a microcontroller an embedded processor? Is an embedded processor a microcontroller? What's the difference between an embedded processor and a microcontroller? Well, today - not much.

Cluster V.

- baud rate
- baud rate generator
- Automatic baud rate detection
- serial port
- operating system
- A2: When Timer 1 is used as the baud rate generator, the baud rates in Modes 1 and 3 are determined by the Timer 1 overflow rate and the value of SMOD (PCON.7 - double speed baud rates) as follows:
- Timer 1 defaults to the baud rate generation for the console port except during execution of any commands that send output to the list port and during the execution of the PWM statement.

- It is interesting to note that the run command leaves timer #1 in auto-baud rate generation mode.
- A typical 8051 contains: - CPU with boolean processor - 5 or 6 interrupts: 2 are external 2 priority levels - 2 or 3 16-bit timer/counters - programmable full-duplex serial port (baud rate provided by one of the timers) - 32 I/O lines (four 8-bit ports) - RAM - ROM/EPROM in some models
- It also has a host of onboard support devices, some members have all of them while others have a subset, the peripherals include: RAM, ROM (mask, OTP, or EEPROM), 2 timers (configurable as timers / counters / comparators / PWM output), watchdog timer, SCI (syncronous serial port), SPI (asynchronous serial port), A/D (8 bit, 8 channel), interrupts.

4. `http://www.aircanada.ca`

- **A Single Long Summary**

  Part I. Top 25 Key Phrases

  - air canada
  - return date
  - departure date
  - news release
  - star alliance
  - air canada fleet
  - fleet fin number
  - fleet air canada
  - fleet air
  - flight attendant
  - canada jazz
  - air canada jazz
  - travel agent

- executive biography
- technical service
- annual information form
- air canada technical service
- air canada technical
- canada technical
- canada technical service
- canada cargo
- air canada cargo
- canada vacation
- air canada vacation
- canada jetz

Part II. Top 25 Key Sentences

- Or, if you wish, you may forward your request and any original unused Air Canada ticket to Air Canada by mail.
- Departure Date and Return Date: A date in the past was chosen.
- Departure Date and Return Date: A Saturday night stay is required and the dates chosen do not meet this condition.
- Departure Date and Return Date: The dates chosen do not meet the minimum stay requirements.
- Departure Date and Return Date: The dates chosen do not meet the maximum stay requirements.
- Departure Date and Return Date: Your date selection does not meet this offers Advance Purchase requirement.
- Brewer acted as a key negotiator in the founding of Star Alliance, of which Air Canada was also a founding member.
- This is the average age of the Air Canada fleet as of this month.

- Air Canada Flight Attendants are ambassadors of the customer experience onboard each and every Air Canada flight.
- Roles of an Air Canada Flight Attendant include Safety professional, caregiver and service provider.
- Air Canada's Flight Attendants make an immediate and lasting positive impression.
- Whether flying on domestic, transborder, or international routes, your child receives extra-special care from Air Canada agents and flight attendants.
- Each year, Air Canada, including Air Canada Jazz, receives thousands of requests for support.
- First among these sub-brands is Air Canada Jazz, Air Canada's primary source for feeder traffic and one of the top four regional carriers in the world.
- Air Canada will allow unaccompanied minors to travel on any itinerary involving Air Canada, Air Canada Jazz and Tier 3 flights.
- It is as easy as asking for an Electronic ticket next time you book a flight with Air Canada or your travel agent.
- If you need any additional information, or have any questions about how to plan for a child travelling on their own, please contact your travel agent or Air Canada.
- Next / we have Air Canada's Technical Services group, which is responsible for the maintenance, repair and overhaul of aircraft.
- Air Canada Jetz is a jet charter service by Air Canada offering premium business service to satisfy the travel needs of corporate clients and professional sports teams.
- Air Canada Jetz is a jet charter service offering a premium business service for corporate groups or sports teams or rock bands.
- Randell is the President and CEO of Air Canada Jazz, which was formed in 2001 with the consolidation of AirBC, Air Ontario, Air Nova and Canadian

Regional Airlines.

- Air Canada's regional airline, Air Canada Jazz, a wholly-owned subsidiary of ACE Aviation Holdings Inc., and its commercial partners serve an additional 48 Canadian and 18 U.S. cities.
- The regional market remains even more bleak with no sign of recovery and Air Canada Jazz recorded an unsustainable operating loss of almost $90 million in the past year.
- We are also moving ahead with the conversion of Air Canada Cargo to a stand-alone subsidiary.
- Starting with the ancillary, non-passenger transportation side of the business, we have Air Canada Cargo.

- **Top 5 Short Cluster Summaries**

  Cluster I.

  - air canada
  - flight attendant
  - travel agent
  - canada jazz
  - air canada jazz
  - Air Canada Flight Attendants are ambassadors of the customer experience onboard each and every Air Canada flight.
  - Roles of an Air Canada Flight Attendant include Safety professional, caregiver and service provider.
  - Air Canada's Flight Attendants make an immediate and lasting positive impression.
  - Whether flying on domestic, transborder, or international routes, your child receives extra-special care from Air Canada agents and flight attendants.
  - It is as easy as asking for an Electronic ticket next time you book a flight with Air Canada or your travel agent.

Cluster II.

- air canada
- star alliance
- canada jazz
- air canada jazz
- air canada jetz
- Brewer acted as a key negotiator in the founding of Star Alliance, of which Air Canada was also a founding member.
- Each year, Air Canada, including Air Canada Jazz, receives thousands of requests for support.
- First among these sub-brands is Air Canada Jazz, Air Canada's primary source for feeder traffic and one of the top four regional carriers in the world.
- Randell is the President and CEO of Air Canada Jazz, which was formed in 2001 with the consolidation of AirBC, Air Ontario, Air Nova and Canadian Regional Airlines.
- As at December 31, 2004, Air Canada employed approximately 32,000 people worldwide including 3,500 at Air Canada Jazz.

Cluster III.

- technical service
- air canada technical service
- air canada technical
- canada technical
- canada technical service
- Air Canada Technical Services, a wholly-owned subsidiary of ACE Aviation Holdings Inc., offers customers worldwide diverse range of technical expertise in the maintenance, repair and overhaul of aircraft, engines, components and various ground and test equipment.

- Air Canada Technical Services also sells ground-handling services to airlines and other customers, as well as training services for mechanics, flight attendants and pilots.
- In March 2003, he assumed the role of President & CEO, Air Canada Technical Services in addition to his responsibilities as Vice President, System Operations Control.
- That includes the possible sale of a significant interest in Air Canada Technical Services and the creation of an Airport Ground Handling Services subsidiary.
- We see these businesses - like Aeroplan and Air Canada Technical Services - as undervalued assets with lots of potential.

Cluster IV.

- air canada vacation
- perfect vacation package
- air canada
- holiday package
- vacation package
- Next is Air Canada Vacations, one of Canada's largest tour operators.
- Find the perfect vacation package to Antigua and book an Air Canada Vacations package today!
- Find the perfect all-inclusive vacation package to Puerto Plata and book an Air Canada Vacations package today!
- Find the perfect vacation package to Aruba and book an Air Canada Vacations package today!
- Find the perfect all-inclusive vacation package to Punta Cana and book an Air Canada Vacations package today!

Cluster V.

- air canada

- canada cargo
- air canada cargo
- canada air
- air canada fleet
- Claude Morin was appointed President & CEO, Air Canada Cargo, a separate limited partnership of ACE Aviation Holdings Inc, in December 2004.
- We are also moving ahead with the conversion of Air Canada Cargo to a stand-alone subsidiary.
- Starting with the ancillary, non-passenger transportation side of the business, we have Air Canada Cargo.
- This is the average age of the Air Canada fleet as of this month.

5. `http://www.cisco.com`

- **A Single Long Summary**

  Part I. Top 25 Key Phrases

  - nortel networks
  - nortel networks limited
  - service provider
  - code division multiple access
  - internet protocol
  - division multiple access
  - local area network
  - wireless local area network
  - nortel networks common share
  - multiple access
  - wireless local area
  - code division

- wireless mesh
- forward looking statement
- wireless mesh network
- networks common share
- universal mobile telecommunications systerm
- wireless mesh network solution
- session initiation protocol
- high speed downlink packet
- network solution
- internet protocol multimedia subsystem
- mesh network
- high performance internet
- internet protocol multimedia

Part II. Top 25 Key Sentences

- Nortel Networks, the Nortel Networks logo and the Globemark are trademarks of Nortel Networks.
- How would you describe the working environment at Nortel Networks? Nortel Networks is an open and transparent company.
- Nortel Networks, the Nortel Networks logo, the Globemark and Business Without Boundaries are trademarks of Nortel Networks.
- About Nortel Networks Nortel Networks is an industry leader and innovator focused on transforming how the world communicates and exchanges information.
- Nortel Networks, the Nortel Networks logo, the Globemark, Business Without Boundaries, Symposium, Alteon and BayStack are trademarks of Nortel Networks.

- The financial results of Nortel Networks Limited ("NNL"), Nortel Networks Corporation's principal operating subsidiary, are fully consolidated into Nortel Networks results.
- General "Associated Companies" means any parent, subsidiary company or affiliate of Nortel Networks Limited.
- We would like to acknowledge the years of contribution and service to the Company and Nortel Networks Limited by Messrs.
- Currie has also been appointed chief financial officer and Sledge, interim controller, of Nortel Networks Limited, the Company's principal operating subsidiary.
- Nortel Networks Limited The financial results of NNL are consolidated into the Company's results.
- Xros' revolutionary technology will play a key role in our strategy to deliver an all-optical Internet, said Clarence Chandran, president, Nortel Networks Service Provider and Carrier Group.
- The Qtera technology brings Nortel Networks solutions closer than ever to the all-optical Internet.
- The Nortel Networks solution will save us $300,000 per year in operational expenses while enabling the entire voice-and-data network to be managed by one person," Adams says.
- Built with a communications infrastructure that features Nortel Networks solutions to support services from the SBC Communications Inc.
- The PacketHop-Nortel Networks solution combines both fixed and mobile Wireless LAN mesh technologies to deliver broadband wireless infrastructure and applications that are highly scalable, survivable and secure.
- As a result of the merger, each outstanding share of Alteon WebSystems common stock was converted into a right to receive 1.83148 Nortel Networks common shares.
- Under terms of the agreement, Alteon WebSystems shareholders will receive a fixed exchange ratio of 1.83148 Nortel Networks common shares for each share of Alteon WebSystems common stock.

- Of the purchase price, an estimated US$705 million will be paid in Nortel Networks common shares at closing on a fully diluted basis.
- As a result of the merger, each outstanding share of Clarify common stock was converted into a right to receive 1.3 Nortel Networks common shares.
- Under the terms of the agreement, Clarify shareholders will receive a fixed exchange ratio of 1.3 Nortel Networks common shares for each share of Clarify common stock.
- Nortel Networks' common shares are listed on the New York, Toronto, Montreal and London stock exchanges.
- As a result, all Nortel Networks common shares will be owned by New Nortel.
- The share exchange will not be taxable to Canadian or United States Nortel Networks common shareholders who hold their shares as capital property.
- A wireless local area network (WLAN) supports remote access to a variety of mobile devices for WakeMed's healthcare professionals - mobile VoIP phones, PDA's, laptops and tablet PCs.
- Nortel's wireless local area networks (LAN) 2200 series portfolio won the security category for its class at SuperComm 2003.

- **Top 5 Short Cluster Summaries**

  Cluster I.

  - nortel networks
  - nortel networks limited
  - service provider
  - bay networks
  - networks common share
  - Nortel Networks, the Nortel Networks logo and the Globemark are trademarks of Nortel Networks.
  - How would you describe the working environment at Nortel Networks? Nortel Networks is an open and transparent company.

- About Nortel Networks Nortel Networks is an industry leader and innovator focused on transforming how the world communicates and exchanges information.
- The financial results of Nortel Networks Limited ("NNL"), Nortel Networks Corporation's principal operating subsidiary, are fully consolidated into Nortel Networks results.
- Nortel Networks' common shares are listed on the New York, Toronto, Montreal and London stock exchanges.

Cluster II.

- wireless networks
- wireless mesh
- wireless solution
- network management
- wireless mesh network
- MeshPlanner is a precise, intuitive planning tool that optimizes wireless mesh network design to reduce planning and implementation cost.
- Nortel sees its Wireless Mesh Network solutions being applied wherever there is a need for communication networks.
- Nortel's Wireless Mesh Network solution is a very versatile technology that fills a special market niche for serving the increasing demand for anywhere, anytime mobile, broadband access.
- Nortel's Wireless Mesh Networks solution introduces the concept of a peer-to-peer mesh topology, with wireless communication between access points.
- Nortel is breaking new ground and leading the wireless broadband evolution by providing our customers and allies with innovative wireless mesh network technology.

Cluster III.

- internet protocol

- local area network
- wireless local area network
- internet protocol multimedia subsystem
- internet protocol multimedia
- A wireless local area network (WLAN) supports remote access to a variety of mobile devices for WakeMed's healthcare professionals - mobile VoIP phones, PDA's, laptops and tablet PCs.
- Nortel's wireless local area networks (LAN) 2200 series portfolio won the security category for its class at SuperComm 2003.
- Wi-Fi - short for wireless fidelity - is the popular term for a high-frequency wireless local area network (WLAN).
- Nortel and Research In Motion announced a joint initiative aimed at delivering secure, full-featured, Session Initiation Protocol-enabled business communications over wireless local area networks.
- Nortel solutions for Internet Protocol (IP) telephony and secure wireless local area networks (LANs) are helping maximize efficiency of communications and collaboration at the Michigan Information Technology Center (MITC), which also serves as Internet2's headquarters.

Cluster IV.

- nortel networks
- network management
- data network
- mesh network
- network solution
- The CDMA2000 1X Wireless Data Network, deployed by Nortel Networks, will enable Multi-Links to significantly boost network capacity to accommodate a greater number of voice calls.
- The Qtera technology brings Nortel Networks solutions closer than ever to the all-optical Internet.

- The Nortel Networks solution will save us $300,000 per year in operational expenses while enabling the entire voice-and-data network to be managed by one person," Adams says.
- Built with a communications infrastructure that features Nortel Networks solutions to support services from the SBC Communications Inc.
- The PacketHop-Nortel Networks solution combines both fixed and mobile Wireless LAN mesh technologies to deliver broadband wireless infrastructure and applications that are highly scalable, survivable and secure.

Cluster V.

- optical networks
- optical networks solutions
- large-scale optical networks
- optical internet
- internet service
- Negative developments associated with Nortel's supply contracts and contract manufacturing agreements, including as a result of using a sole supplier for a key component of certain optical networks solutions
- Xros' revolutionary silicon-based micro-mirror technology will allow data to be switched through large-scale optical networks entirely in the form of light.
- StorageXtend was tested over WilTel's nationwide SONET network to ensure interoperability with EMC and Nortel Networks Business Continuity Over Optical Networks solution.
- With a clear recognition of growing customer need for more cost-effective and simplified MAN and WAN networking options, Nortel and EMC have jointly developed a set of Business Continuity over Optical Networks solutions, to enable enterprise customers to easily deploy high availability data replication for their storage area network.

6. `http://www.oracle.com`

- **A Single Long Summary**

  Part I. Top 25 Key Phrases

  - service oriented architecture windows
  - security service oriented architecture
  - architecture windows server
  - applications technology product
  - management applications technology
  - management application
  - windows server system
  - management applications technology product
  - architecture windows server system
  - enterprise management application
  - applications technology
  - enterprise management applications technology
  - peoplesoft enterprise
  - technology products a z
  - enterprise management
  - technology product
  - e business suite
  - tools enterprise management
  - oracle e business suite
  - windows server system technologies
  - jd edwards
  - oracle aces techblast newsletter
  - developer tools enterprise management
  - products a z

- tools enterprise management application

Part II. Top 25 Key Sentences

- PeopleSoft Enterprise HelpDesk PeopleSoft Enterprise HelpDesk enhances the overall speed and quality of internal support operations by optimizing the efforts of your help desk staff and providing comprehensive process automation.
- Oracle Enterprise Manager Oracle Enterprise Manager with Oracle Grid Control provides a single, integrated interface for administering and monitoring applications and systems in an Oracle Grid.
- We will still continue to deliver tax, legal, and regulatory updates for six-years for the PeopleSoft Enterprise and JD Edwards EnterpriseOne applications.
- JD Edwards World belongs to the Oracle Applications product line, which also includes PeopleSoft Enterprise, JD Edwards EnterpriseOne, and the Oracle E-Business Suite.
- Oracle E-Business Suite On Demand Oracle's end-to-end business applications for CRM, ERP, SCM, and more, without the demands of monitoring and maintenance.
- Oracle E-Business Suite is a fully integrated, comprehensive suite of business applications for the enterprise.
- Oracle E-Business Suite provides businesses the functional best practices and industry-specific capabilities they need to adapt to change and compete more effectively.
- The Oracle E-Business Suite family of Marketing Applications provides true information-driven marketing, enabling you to plan, execute, and analyze marketing campaigns and complex trade promotions.
- Implement one or several application families – or implement the complete Oracle E-Business Suite for the fastest way to high-quality enterprise information.

- Oracle E-Business Suite family of Order Management applications streamline and automate the entire sales order management process, from order promising and order capture to transportation and shipment.
- Implement one or several application families–or implement the complete Oracle E-Business Suite for the fastest way to high-quality enterprise information.
- Oracle Advanced Procurement is a key component of the Oracle E-Business Suite.
- Oracle's solution includes a Compliance package, an RFID pilot kit, and integrated support in Oracle E-Business Suite and Oracle Application Server.
- Oracle cMRO touches 22 applications in the Oracle E-Business Suite to provide a comprehensive air transportation maintenance and A&D MRO service solution.
- 7-Eleven "We bought the entire Oracle E-Business Suite because it was pre-integrated, pre-tested, and pre-defined.
- In fact Oracle builds security into all of its products–not just into Oracle Database but also Oracle Application Server, Oracle Collaboration Suite, and Oracle E-Business Suite.
- These are a few examples of the advanced support technologies available to Oracle E-Business Suite and Oracle technology customers.
- Oracle E-Business Suite Customers Applications Customers [spacer.gif] [spacer.gif] Oracle E-Business Suite offers a complete set of applications capable of automating any daily business process.
- We also chose Oracle Fusion Middleware and the BPEL Process Manager in order to deploy and extend our next generation of Oracle E-Business Suite Applications."
- Plugged In [See a sample] News, events, and support information for E-Business Suite, PeopleSoft Enterprise, JD Edwards EnterpriseOne, JD Edwards World, and Siebel customers.
- Oracle Applications Oracle Applications, including the Oracle E-Business Suite, PeopleSoft Enterprise, JD Edwards EnterpriseOne and JD Edwards

World, enable information-driven business processes that connect and automate an organization.

- Oracle's certification of PeopleSoft, Oracle E-Business Suite and JD Edwards EnterpriseOne applications with Oracle Fusion Middleware is expected to benefit customers both immediately and in the long-term.
- JD Edwards EnterpriseOne and JD Edwards World On Demand High availability and continuous availability solutions specifically designed to support the JD Edwards product family.

- **Top 5 Short Cluster Summaries**

  Cluster I.

  - collaboration suite
  - oracle collaboration
  - oracle collaboration suite
  - oracle application
  - oracle product
  - About Oracle Collaboration Suite Oracle Collaboration Suite is the first enterprise-class collaboration product built on the right architecture: Oracle Database and Oracle Application Server.
  - Especially attractive to us is Oracle Collaboration Suite's ease of use and the security offered by Oracle applications and database products.
  - Oracle Collaboration Suite 10g Content Services is a robust platform priced and designed for enterprise-wide records management, business process automation and integrated file and document management.
  - Oracle Collaboration Suite On Demand is a complete management service for messaging, calendaring, file sharing, and real-time communications.
  - Oracle Database, Oracle Application Server, and Oracle Collaboration Suite address all your information integration needs so you can make better business decisions faster.

Cluster II.

- oracle fusion
- fusion middleware
- oracle fusion middleware
- oracle application
- application server
- Oracle Fusion Middleware has a range of interoperability with Microsoft .Net and Office.
- Oracle Fusion Middleware is built on the industry's most complete, J2EE and open-standards-based integration infrastructure.
- Oracle Portal is a member of the Oracle Fusion Middleware family of products, which bring greater agility, better decision-making, and reduced cost and risk to diverse IT environments today.
- Oracle Fusion Middleware is a family of standards-based, customer-proven products that includes Oracle Application Server and related tools and options, Oracle Collaboration Suite, and Oracle Data Hubs.
- The underlying information infrastructure of Oracle Fusion Middleware maximizes the benefits of information-driven applications through the combination of the Oracle Database and Oracle Applications.

Cluster III.

- application server
- oracle database
- oracle application
- oracle application server
- oracle product
- Oracle Database, Oracle Application Server, and Oracle Collaboration Suite address all your information integration needs so you can make better business decisions faster.

- Oracle offers the only complete, integrated software stack–Oracle Database 10g and Oracle Application Server 10g–engineered to work together to enable grid computing.
- In fact Oracle builds security into all of its products–not just into Oracle Database but also Oracle Application Server, Oracle Collaboration Suite, and Oracle E-Business Suite.
- Finally, Oracle technology, including Oracle Database and Oracle Application Server, integrates all your employee information for an accurate view of training levels, benefits, and status across your organization.
- The portal architecture is provided through Oracle Application Server 10g and includes the maximum set of functionality, including Oracle Discoverer, Reports, and Forms.

Cluster IV.

- e business suite
- oracle e business
- oracle e business suite
- business process
- business intelligence
- Oracle E-Business Suite On Demand Oracle's end-to-end business applications for CRM, ERP, SCM, and more, without the demands of monitoring and maintenance.
- Oracle E-Business Suite is a fully integrated, comprehensive suite of business applications for the enterprise.
- Oracle E-Business Suite provides businesses the functional best practices and industry-specific capabilities they need to adapt to change and compete more effectively.
- The Oracle E-Business Suite family of Marketing Applications provides true information-driven marketing, enabling you to plan, execute, and analyze marketing campaigns and complex trade promotions.

- Oracle Advanced Procurement is a key component of the Oracle E-Business Suite.

Cluster V.

- management service
- customer relationship
- relationship management
- customer relationship management
- management solution
- All Campus Solutions Product Modules Customer Relationship Management Increase revenues and drive customer satisfaction and loyalty through Sales, Marketing, and Service effectiveness.
- Oracle's PeopleSoft Customer Relationship Management (CRM) Warehouse captures all customer-related data into a single repository and combines it with complex analysis of your sales, marketing, and service initiatives.
- Customer Relationship Management Metrics and KPIs PeopleSoft Customer Scorecard provides more than 25 predefined metrics and KPIs that provide quick and simplified views into the complete customer lifecycle.