

BİLAL GENÇ

CONTENT ANALYSIS OF UN-STRUCTURED
LOCAL INTERNET NEWS WEBSITES

BİLAL GENÇ

September, 2011

ÇANKAYA UNIVERSITY

CONTENT ANALYSIS OF UN-STRUCTURED
LOCAL INTERNET NEWS WEBSITES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ÇANKAYA UNIVERSITY

BY

BİLAL GENÇ


IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2011

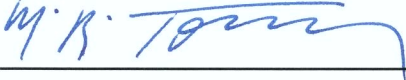
Title of the Thesis: **Content Analysis of Un-Structured Local Internet News Websites**

Submitted by **Bilal GENÇ**

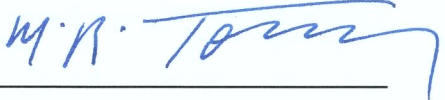
Approval of the Graduate School of Natural and Applied Sciences, Çankaya University


Prof. Dr. Taner ALTUNOK
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.


Prof. Dr. Mehmet R. TOLUN
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.


Prof. Dr. Mehmet R. TOLUN
Supervisor

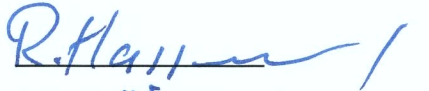
Examination Date : 12.09.2011

Examining Committee Members:

Prof. Dr. Mehmet R. TOLUN (Çankaya Univ.)



Asst. Prof. Dr. Reza HASSANPOUR (Çankaya Univ.)



Asst. Prof. Dr. Kasım ÖZTOPRAK (KTO Karatay Univ.)



STATEMENT OF NON-PLAGIARISM

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : Bilal Genç

Signature : 

Date : 12.09.2011

ABSTRACT

CONTENT ANALYSIS OF UN-STRUCTURED LOCAL INTERNET NEWS WEBSITES

GENÇ, Bilal

M.Sc., Department of Computer Engineering

Supervisor: Prof. Dr. Mehmet R. Tolun

September 2011, 43 pages

In today's world data is a real power. In order to get advantage of the data power, analysis of the data is very important. Social incidents have been analyzed for more than a century. In order to understand social incidents better, data has big importance. Localized social analysis can be easily done by analyzing the local Internet content. In this study methods for analysis of the local news websites are discussed. On the other hand, a solution is introduced to overcome problems of un-structured website designs such as Turkish character set problems, non standard development techniques, unrelated contents such as advertisements and comments. An algorithm and code was developed to filter and index news website content. As a result code was implemented in a website and proved to be running.

Keywords: Content Analysis , Computer Aided Text Analysis, Website Analysis

ÖZ

STANDART DIŐI YEREL INTERNET HABER SİTELERİNİN İÇERİK ANALİZİ

GENÇ, Bilal

Yükseklisans, Bilgisayar Mühendisliđi Anabilim Dalı

Tez Yöneticisi: Prof. Dr.Mehmet R. Tolun

Eylül 2010, 43 sayfa

Günümüz dünyasında veri gerçek güçtür. Veriden en yüksek avantajı sağlamak için analizinin yapılması çok önemlidir. Sosyal olaylar yaklaşık bir asırdan fazla zamandır analiz edilmektedir. Sosyal olayların daha iyi anlaşılabilmesi için yerel bilgi çok büyük önem arz etmektedir. Bölgesel sosyal analiz, yerel Internet içeriklerinin analizi ile rahatlıkla yapılabilir. Bu çalışmada yerel Internet haber sitelerinin analizi ile ilgili metodlar tartışılmıştır. Ayrıca, Türkiye’de bulunan haber web sitelerinin analizine ilişkin Türkçe karakter seti kaynaklı problemler, standart olmayan geliştirme, portal yapıları, site içeriđiyle ilgili olmayan reklam ve yorum içeriklerinden kaynaklı problemlere çözüm sağlanmıştır. Web sitesinin içeriđini filtreleyen ve indeksleyen bir algoritma oluşturulmuştur ve kod geliştirilmiştir. Sonuç olarak kod bir web sitesinde uygulanmış ve çalıştığı ispat edilmiştir.

Anahtar Kelimeler: İçerik Analizi, Bilgisayar Destekli İçerik Analizi, Web Sayfası Analizi

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor Prof. Dr. Mehmet R. Tolun and Ass. Prof. Dr. Kasım Öztoprak for their guidance, advices, criticism, encouragements, also my wife Burçin Genç and my son Burak Genç for forgiving me for using our precious time throughout the research. Also I would like to thank M.Fatih Soydan General Manager and owner of Usishi Company for supporting me technically and giving a chance to make a study on their core business area and E.Çınar Çolakođlu for supporting me during the writing of this thesis.

TABLE OF CONTENTS

STATEMENT OF NON-PLAGIARISM	iii
ABSTRACT	iv
ÖZ	v
ACKNOWLEDGEMENTS	vi
TABLE OF CONTENTS	vii
LIST OF TABLES	ix
LIST OF FIGURES	x
LIST OF SYMBOLS AND ABBREVIATIONS	xi
1. INTRODUCTION	1
1.1 Background	1
1.2 Motivation	3
1.3 Objective	4
1.4 Thesis Outline	4
2. THEORETICAL BACKGROUND	5
2.1 Definiton of Content Analysis	5
2.2 Literature Review	6
2.3 Uses of Content Analysis	7
2.4 The Process of a Content Analysis	8
2.5 Content Analysis In WWW	10
3. QUANTITATIVE AND QUALITATIVE ANALYSIS	14
3.1 Qualitative and Quantitative Analysis	14
3.2 Measuring Context	15
3.3 Conducting Contextual Content Analysis (McTavish's Method)	16

3.4	Procedures of Qualitative Content Analysis	20
3.4.1	Inductive category development	20
3.4.2	Deductive category development	21
4.	CATA – COMPUTER AIDED TEXT ANALYSIS.....	24
4.1	Coding Methods in Text Analysis	24
4.2	CATA Tools.....	25
4.3	Quantitative and Qualitative CATA Tools	26
5.	PROPOSED SOLUTION FOR ANALYSIS OF TURKISH NEWS WEBSITES	27
5.1	Challenges.....	27
5.2	Technologies Used.....	28
5.3	Conceptual Design	28
5.3.1	Crawler	29
5.3.2	Text search.....	29
5.3.3	Page analyzer	29
5.3.4	Database.....	30
5.3.5	Services.....	30
6.	EXPERIMENTAL STUDIES AND RESULTS.....	31
6.1	Crawler.....	31
6.2	Infrastructure Information.....	32
6.3	Content Deduction	37
6.4	Results.....	39
7.	CONCLUSION	42
7.1	Future Work.....	43
	REFERENCES.....	R1
	APPENDIX.....	A1
	A. CODE.....	A1

LIST OF TABLES

TABLES

Table 2.1	Uses of Content Analysis by Purpose, Communication Element, and Question	8
Table 3.1	Illustrative Conceptual Categories and Words Four Context Categories.....	16
Table 3.2	Accumulation of C-Scores for the Illustrative Sentence	19
Table 4.1	CATA Programs.....	26
Table 6.1	Web Site List Schema	34
Table 6.2	Web Page List and Categories	34
Table 6.3	Content Type	35
Table 6.4	City List.....	35
Table 6.5	Site Group List	36
Table 6.6	Educational Statistics	36
Table 6.7	Regex Removal List	37
Table 6.8	Code Removal List.....	38
Table 6.9	Processed and Unprocessed Content.....	41

LIST OF FIGURES

FIGURES

Figure 3.1	Step Model for Inductive Category Development	21
Figure 3.2	Step Model for Deductive Category Application	22
Figure 5.1	Conceptual Design for the Solution.....	29
Figure 6.1	Flowchart of Page Analysis Class	33
Figure 6.2	Sample Web Page	40
Figure 6.3	Content Title and Information in the Database.....	40

LIST OF SYMBOLS AND ABBREVIATIONS

CATA	Computer Aided Text Analysis
KWIC	Keyword in Context
C- Score	Contextual Score
E-Score	Idea/Emphasis Score
MCCA	Minnesota Contextual Content Analysis
SQL	Structured Query Language

CHAPTER 1

INTRODUCTION

In this chapter, fundamentals of content analysis are discussed. First definitions and background is provided. Then an introduction to computerized methods is made and motivation for the thesis and the objective is given. Finally outline of the thesis is discussed.

1.1 Background

Content analysis has been an important topic for social sciences starting from 1960's. First hypothesis was created before 1930's, but in 1931 a much more accurate methodology was developed by Alfred R Lindesmith which became known as a content analysis technique [1]. The Constant Comparative Method of Qualitative Analysis" first published in 1964-65 by Glaser and Strauss and referred to their adaptation of it as "Grounded Theory" [1]. In this theory "Key Word in Context" (KWIC) was introduced. KWIC is the content analysis method enables systematically identify word properties, such as the frequencies of most used keywords by locating the more important structures of its communication content. In 1975 David Robertson created a coding frame for categorization and comparison used in political party analysis. it has been developed further by Manifesto Research Group [2] in 1979.

In 1980's public - media relations has realized the power of content analysis and data from circulation, readership, number of viewers and listeners have been used for content analysis. Famous futurist John Naisbutt used these analysis in his publication Megatrends in 1982 [3]. In 1980's Holsti's [4] studies moved forward

by Krippendorff's book of bible in Content Analysis (Content analysis: An introduction to its methodology). In this book Krippendorff introduced units of analysis as the sampling unit, the recording unit, and the context unit [5].

According to Dr. Klaus Krippendorff (1980 and 2004), six questions must be addressed in every content analysis [5]:

1. Which data are analyzed?
2. How are they defined?
3. What is the population from which they are drawn?
4. What is the context relative to which the data are analyzed?
5. What are the boundaries of the analysis?
6. What is the target of the inferences?

According to Zipf's law, the assumption is that words and phrases mentioned most often are those reflecting important concerns in every communication. Therefore, quantitative content analysis starts with word frequencies, space measurements (column centimeters/inches in the case of newspapers), time counts (for radio and television time) and keyword frequencies [6]. However, content analysis extends far beyond plain word counts. Synonyms and homonyms can be isolated in accordance to linguistic properties of a language.

Qualitatively, content analysis can involve any kind of analysis where communication content (speech, written text, interviews, images ...) is categorized and classified. In its beginnings, using the first newspapers at the end of 19th century, analysis was done manually by measuring the number of lines and amount of space given a subject. With the rise of common computing facilities like PCs, computer-based methods of analysis are growing in popularity. Answers to open ended questions, newspaper articles, political party manifestoes, medical records or systematic observations in experiments can all be subject to systematic analysis of textual data. By having contents of communication available in form of machine readable texts, the input is analyzed for frequencies and coded into categories for building up inferences. Robert Philip Weber [7] notes: "To make valid inferences

from the text, it is important that the classification procedure be reliable in the sense of being consistent: Different people should code the same text in the same way". The validity, inter-coder reliability and intra-coder reliability are subject to intense methodological research efforts over long years [8]. Dermot McKeone [9] has highlighted the difference between prescriptive analysis and open analysis. In prescriptive analysis, the context is a closely-defined set of communication parameters (e.g. specific messages, subject matter); open analysis identifies the dominant messages and subject matter within the text.

A further step in analysis is the distinction between dictionary-based (quantitative) approaches and qualitative approaches. Dictionary-based approaches set up a list of categories derived from the frequency list of words and control the distribution of words and their respective categories over the texts. While methods in quantitative content analysis in this way transform observations of found categories into quantitative statistical data, the qualitative content analysis focuses more on the intentionality and its implications.

Dr. Kimberly A. Neuendorf [10] suggests that when human coders are used in content analysis, reliability translates to intercoder reliability or "the amount of agreement or correspondence among two or more coders.". The website of Dr. Neuendorf's Content Analysis Guidebook [11] offers further information about assessment of reliability, reliability coefficients, calculations of intercoder reliability, resources, bibliographies on content analysis, an exhaustive list of message archives and other downloadable CATA programs.

1.2 Motivation

Today measurement of the audience attitude and emotions is very important. In the Middle East Region civil movements have been organized by local digital media and it has become very important to analyze. In Turkey a lot of local news websites have been followed by people and their importance is being increased day by day. Although a number of studies have been achieved for content analysis of the web sites, it has lack of a localized approach. Localization means to analyze according to the language, social background and understanding. Another challenge is to have a

methodology for unstructured websites those have been created by local web developers which have no coding discipline. In this study motivation is to achieve a reliable and localized approach to content analysis of local news websites.

1.3 Objective

The two objectives of this thesis are as follows: 1. Creating a model for unstructured web sites analysis. 2. Software as a Proof of Concept for content analysis of the websites

1.4 Thesis Outline

Thesis consists of the following parts:

Theoretical definitions of content analysis are given in chapter 2. The chapter is intended as introduction for basic concepts and methods of content analysis. The methods of website analysis are discussed as well.

Chapter 3 gives further information about content analysis. In this chapter Qualitative and Quantitative Analysis methods are discussed.

Chapter 4 introduces Computer Aided Text Analysis and software.

Chapter 5 elaborates a method created for the unstructured news websites analysis in Turkey Sample workflows and building blocks for the concept will be introduced

Chapter 6 introduces an application developed according to the proposed concept and analysis algorithms. Also sample measures and results are given.

Chapter 7, the thesis ends up with the summary and conclusions part including important conclusions from this study. Finally, the future possible work related with this study is given.

CHAPTER 2

THEORETICAL BACKGROUND

This chapter consists of theoretical definitions of content analysis and its web based approach.

2.1 Definiton of Content Analysis

Content Analysis defined by a lot of social science experts, to academicians. More basic definitons tell that it is a methodology in the social sciences for studying the content of communication. Dr. Farooq Joubish, content analysis is considered a scholarly methodology in the humanities by which texts are studied as to authorship, authenticity, or meaning; including philology, hermeneutics, and semiotics [12].

According to Ole Holsti [4] content analysis is "any technique for making inferences by objectively and systematically identifying specified characteristics of messages. More systematic one is introduced by Kimberly A. Neuendorf [10] offering six criteria for content analysis: "Content analysis is a summarizing, quantitative analysis of messages that relies on the scientific method (including attention to objectivity, intersubjectivity, a priori design, reliability, validity, generalisability, replicability, and hypothesis testing) and is not limited as to the types of variables that may be measured or the context in which the messages are created or presented."

2.2 Literature Review

Since the publication of early treatises in the middle of the 20th century by Berelson in 1952 [13] and Berelson & Lazarsfeld, 1948 [14], content analysis has been adopted by a wide range of academic disciplines including communications, psychology, sociology, organizational theory, and political science. It is appropriate that with the much-heralded rise of the information society, researchers have continued to uncover new data and research questions that require the rigorous analysis of message content. Content analyses have examined a wide variety of media messages including newspaper articles, television programs, one-on-one conversations, and advertising images. Although the research questions that may be addressed through content analysis are almost limitless, Holsti provides a useful three-part typology [4]. First, researchers may focus on message content and meaning, such as efforts to assess the propaganda content of messages or to make intermedia comparisons of messages. Second, researchers may examine the antecedents of messages. For example, psychologists may analyze the writings and conversations of patients to evaluate their state of mind. Finally, researchers may examine message effects, such as the voluminous research on the behavioral consequences of viewing television violence. Although the structure of a content analytic study depends on the specific research question being addressed, the outlines of the procedures required to develop valid and reliable measures and inferences are well established. With a research question or hypothesis in hand, the researcher first defines the population of messages addressed by the research question and selects a sample from that population.

Krippendorff [5] differentiates between three units: the sampling unit, the recording unit, and the context unit. The sampling unit encompasses the whole independent message used for the basis of sampling. Recording units are analytically separable components of the message that will be independently attached to content categories. In an analysis of newspaper articles, for example, an article can be divided into a number of syntactically defined recording units such as the entire article, paragraphs, sentences, or individual words. It is normally desirable to select the smallest recording unit practical because determining the nature of the content is

easier for coders and the measures of smaller units can always be aggregated to measure larger units. Generally, the choice of recording units involves a trade-off between reliability and validity. A researcher may define recording units employing a more refined understanding of message structure such as the propositional structure of messages or underlying themes. Although such divisions may be more analytically useful, they require a deeper understanding of the underlying language, making the analysis more labor intensive and less reliable [5]. To address the fact that message meanings are contextual, researchers also define a context unit, the immediate environment in which the message is received. An entire newspaper, for example, could be the context unit for a news article. The third and most important element of content analysis is the development of a categorization scheme by which messages can be validly and reliably classified. Because of the inherently ambiguous and complex nature of messages and the variety of research interests, many different categorization schemes are employed. While some researchers argue that categories should focus exclusively on the manifest components of messages (Berelson, 1952), others often attempt to assess latent characteristics of communications employing judgmental scales. Some standard category schemes include the following: (a) the gross characteristics of messages such as length, (b) frequency counts of how often an idea or subject occurs in a message, (c) assessments of what message components are emphasized, (d) the relationship between message components, and (e) the number of qualifications made and associations expressed toward an idea or subject Krippendorff [5] Fourth, the researcher trains coders to code the sample of messages according to the categorization scheme developed and checks intercoder reliability. Finally, the researcher analyzes and interprets the data collected

2.3 Uses of Content Analysis

Ole Holsti [4] groups 15 uses of content analysis into three basic categories:

- make inferences about the antecedents of a communication
- describe and make inferences about characteristics of a communication
- make inferences about the effects of a communication.

The following table shows fifteen uses of content analysis in terms of their general purpose, element of the communication paradigm to which they apply, and the general question they are intended to answer.

Table 2.1: Uses of Content Analysis by Purpose, Communication Element, and Question

Purpose	Element	Question	Use
Make inferences about the antecedents of communications	Source	Who?	Answer questions of disputed authorship (authorship analysis)
	Encoding process	Why?	Secure political & military intelligence Analyse traits of individuals Infer cultural aspects & change Provide legal & evaluative evidence
Describe & make inferences about the characteristics of communications	Channel	How?	Analyse techniques of persuasion Analyse style
	Message	What?	Describe trends in communication content Relate known characteristics of sources to messages they produce Compare communication content to standards
	Recipient	To whom?	Relate known characteristics of audiences to messages produced for them Describe patterns of communication
Make inferences about the consequences of communications	Decoding process	With what effect?	Measure readability Analyse the flow of information Assess responses to communications

2.4 The Process of a Content Analysis

According to Dr. Klaus Krippendorff [5,10] six questions must be addressed in every content analysis:

1. Which data are analyzed?
2. How are they defined?
3. What is the population from which they are drawn?
4. What is the context relative to which the data are analyzed?
5. What are the boundaries of the analysis?
6. What is the target of the inferences?

The assumption is that words and phrases mentioned most often are those reflecting important concerns in every communication. Therefore, quantitative content analysis starts with word frequencies, space measurements (column centimeters/inches in the case of newspapers), time counts (for radio and television time) and keyword frequencies. However, content analysis extends far beyond plain word counts, e.g. with Keyword In Context routines words can be analyzed in their specific context to be disambiguated. Synonyms and homonyms can be isolated in accordance to linguistic properties of a language.

Qualitatively, content analysis can involve any kind of analysis where communication content (speech, written text, interviews, images ...) is categorised and classified. In its beginnings, using the first newspapers at the end of 19th century, analysis was done manually by measuring the number of lines and amount of space given a subject. With the rise of common computing facilities like PCs, computer-based methods of analysis are growing in popularity. Answers to open ended questions, newspaper articles, political party manifestoes, medical records or systematic observations in experiments can all be subject to systematic analysis of textual data. By having contents of communication available in form of machine readable texts, the input is analyzed for frequencies and coded into categories for building up inferences. Robert Philip Weber (1990) [7] notes: "To make valid inferences from the text, it is important that the classification procedure be reliable in the sense of being consistent: Different people should code the same text in the same way" [8]. The validity, inter-coder reliability and intra-coder reliability are subject to intense methodological research efforts over long years

One more distinction is between the manifest contents (of communication) and its latent meaning. "Manifest" describes what (an author or speaker) definitely has written, while latent meaning describes what an author intended to say/write. Normally, content analysis can only be applied on manifest content; that is, the words, sentences, or texts themselves, rather than their meanings.

Dermot McKeone [9] has highlighted the difference between prescriptive analysis and open analysis. In prescriptive analysis, the context is a closely-defined set of

communication parameters (e.g. specific messages, subject matter); open analysis identifies the dominant messages and subject matter within the text.

A further step in analysis is the distinction between dictionary-based (quantitative) approaches and qualitative approaches. Dictionary-based approaches set up a list of categories derived from the frequency list of words and control the distribution of words and their respective categories over the texts. While methods in quantitative content analysis in this way transform observations of found categories into quantitative statistical data, the qualitative content analysis focuses more on the intentionality and its implications.

2.5 Content Analysis In WWW

In the past two decades, Internet-related research has been moving from early descriptive studies about the medium itself the Internet's characteristics to a higher level, focusing on Web users and social effects. Kopper and colleagues, for instance, have identified seven perspectives that current Internet research has been pursuing, including analyses of product, users, quality, social context, market, occupational changes, and experimental projects [15]. Meanwhile, Stempel and Stewart also point out methodological challenges that communication researchers had to deal with, especially in audience research and content analysis [16]. In contrast to the traditional media studies, where quantitative research has prevailed for many years, Kim and Weaver discover that early Internet research used non-quantitative methods more frequently than quantitative methods [16]. Some assumed that difficulty in collecting online data was one of the main constraints preventing quantitative studies; in particular, they point to sampling problems in content analysis.

While efforts have been made to explore to take advantage of the Internet to conduct online surveys, content analysis continuous to face many challenges in measuring the hyper textual and interactive Web content, as well as problems of sampling, unitization, and coding.

The mix of various media characteristics makes Internet content analysis fairly complex. Stempel and Stewart [16] concern was how such complexity affects generalizability and representativeness regarding the Web content analysis. McMillan [18] concluded that five major issues exist when examining Web content:

1. how to identify the units to be sampled?
2. how to collect data for cross-coder tests when the Web changes rapidly?
3. how to solve copyright issues if researchers download Web pages for analysis?
4. how to standardize units of analysis given the multimedia features of the Web?
5. how to check inter-coder reliability?

Similarly, Weare and Lin examined the potential methodological issues of content analysis and identified problems existing in the processes of sampling, unitization, categorization, and coding [17]. In particular, McMillan [18] recommended that researchers investigate the validity of multiple sampling methods on the Web. The goal of sampling is to generate a manageable subset of data from a large population or a sampling frame to represent this population. An ideal sample is a tradeoff between the ease for study and the representativeness of the population. Thus, content analysts should determine how to define a tangible sampling frame, how to draw a representative sample from the sampling frame, and how large the sample size must be to be not only effective but also efficient.

On one hand, the amount of information on the Web is enormous and expands at an exponential rate. On the other hand, the decentralized nature of cyberspace allows any Web user to create and transmit various forms of information anytime from anywhere. The anonymity makes it even harder to estimate the sampling frames for content analytic research.

In practice, content analysts commonly used online search engines and available directories for their sampling frames. For instance, Joseph R. Dominick [19] located 500 personal home pages via Yahoo! Directory; Mary Paul [20] found 64 disaster

relief home pages by using several online search engines; and Liu et al [21] analyzed business Web sites by using the Fortune 500 companies index. However, using research engines and assorted directories was still problematic because Web sites emerge and recede too rapidly to be traced, choosing key words for searching is a tricky business, and even the most sophisticated search engines can find only a small amount of information online. Some claimed the size of the “invisible” data to search engines was 500 times the content that could be searched. Sample size is another issue involved in sampling of the Internet [22], which, however, has drawn little attention from Internet content analysts. Part of reason might be the difficulty of testing sample size effectiveness because researchers are not even aware of the size of sampling frames, including both visible and invisible data. Thus, the invisibility of Web content toward search engines essentially leads to the fact that researcher can merely draw a convenient sample using search engines.

Questions about sampling method on the Internet might leave researchers feeling hopeless. But before being pessimistic, researchers need to notice not all sampling frames are indefinable on Web, depending on units of analysis. For instance, analyses of a group of Web sites may initially be plagued by troubles in defining the group, because those Web sites are essentially independent or isolated in the cyber world. Researchers cannot detect them unless they are linked together. Search engines have established a form of linkage to gather these Web sites, partially not comprehensively, which causes the problem of defining sampling frames.

On the contrary, longitudinal research designs have a relatively explicit sampling frame, such as Li’s study of newspapers’ Web pages design [23], Massey and Chang’s analysis of Asian Web newspapers [24], and Cassidy’s comparison of Web-only news sites and daily paper sites [25]. Their units of analysis are articles or Web pages within a specific site rather than isolated Web sites. Since articles and Web pages are connected to a Web site, they are all visible. Thus researchers should be able to estimate the sampling frames the overall articles or Web pages within a certain time frame. Then, whether a sample size is representative should have

become researchers' concern, which unfortunately has not been addressed clearly or convincingly.

Although the sampling methods varied depending on their specific research questions in these studies, such variation suggest that content analysts needed a sampling guideline for examining the Web or at least some assumptions about sampling need testing with Web content.

CHAPTER 3

QUANTITATIVE AND QUALITATIVE ANALYSIS

In this chapter, Content Analysis Methods will be discussed and computerized methods will be introduced.

3.1 Qualitative and Quantitative Analysis

During the earlier days of content analysis declared that in 1950's first attempts for enriching the methodology of content analysis have been started. The quantitative requirement for content analysis has stressed the importance and relevance to analysis of the frequency with which words or themes appear in a body of texts. It has been clearly demonstrated by Holsti [4] that restricting content analysis to frequency counts presents theoretical and practical problems and that measures other than frequency may be useful. Besides it has been clearly seen that qualitative analysis should be an act of interpretation which is not based on measurement of frequencies of occurrences or statistical significance tests. On the other hand the rigidity of quantitative content analysis that based purely on counts, frequencies etc is more artificial and studies should include non-numerical procedures at various stages in the research such as the initial selection of categories, or check validity of the quantitative results after coding has been performed. Besides quantitative results may highlight qualitative aspects of the text which might have escaped researchers attention. In 1969, Holsti declared that " the content analysts should use both qualitative and quantitative methods to supplement each other: by moving back and forth between these approaches that the investigator is most likely to gain insight into meaning of his data [4]. According to McTavish [26] " The approach makes quantitative distinctions between texts varying in both the pattern of emphasis upon different sets of ideas and in the context or social perspective from which these ideas are addressed. Scores are used to describe comparative patterns of meaning in

textual data, generate traditional statistical analyses with other non-textual variables, and aid in organizing and focusing further qualitative analysis. “

To aid in interpreting contextual information in these profiles, McTavish offered two scores basically: 1. E- Score (emphasis score) 2. C- Score (Contextual score). E Score is much more quantitative and enables investigator to examine the over- and under-emphasis on idea categories relative to the norm of expected category usage. Broader concepts and themes in a text can be identified from scores for sets of related categories. The C- Score is rather Qualitative and distinctions between texts can be made in terms of the overall profile of emphasis on idea categories uses four general "marker" contexts we call "traditional", "practical", "emotional", and "analytic". Each marker context is an experimental, empirically-derived profile of relative emphasis on each idea category, which characterizes the perspective typical of a general social or institutional context. As a set, the four contextual markers serve as dimensions to define a social context space and contextual scores (called C-scores) are computed. For example, an investigator can realistically examine transcribed conversational interviews on a topic for a large representative sample of cases. Quantitative scores can help guide comparative, qualitative analysis of social meanings in textual data, adding depth and anchoring to quantitative causal analyses as well.

3.2 Measuring Context

Main purpose of measuring is to understand general social meaning. According to McTavish's frame work of contextual dimensions, there are four general contexts: (a) traditional, (b) practical, (c) emotional, and (d) analytic [26]

Each of the four contexts incorporates a general idea of societal activity and represents a different framework within which specific concepts can emerge:

Each of the four contexts incorporates a general idea of societal activity and represents a different framework within which specific concepts can emerge:

- a) Traditional Context: A normative perspective on the social situation predominates and the situation is defined in terms of standards, rules and codes which guide social behavior.
- b) Practical Context: A pragmatic perspective of the social situation predominates and behavior is directed toward the rational achievement of goals.
- c) Emotional Context: An affective perspective predominates and the situation is defined in terms of expressions of emotion (both positive and negative), and maximizing individual involvement, personal concern and comfort.
- d) Analytic Context: An intellectual perspective predominates and the situation is defined in objective terms.

Table 2 schematically illustrates the way in which word groupings in a conceptual dictionary can reflect idea categories and certain idea categories may be emphasized more heavily in certain social contexts.

Table 3.1: Illustrative Conceptual Categories and Words Four Context Categories

Context	Category	Typical word or phrase
Traditional	Guide	should, ought, guard
	Structural Roles	mighty, military
	Prohibit	restrict, watch
	Ideals	stability, honesty
Practical	Activity	walk, buy, sell
	Merchandise	product, spend
	Strive	maintenance, development
	Organization	management, office, factory, retail
Emotional	Happy	friendly, wonderful
	Pleasure	gladness, refreshment
	Expression Arena	museum, music
	Self-other	respond, wish
Analytic	Differentiate	analysis, analytic
	Relevant	solution, signify
	Similarity	alike, comparison
	Scholarly Nouns	library, university, science

3.3 Conducting Contextual Content Analysis (McTavish’s Method)

According to McTavish conceptual content analysis focuses on ideas in text [26]. Contextual content analysis specifies perspective to those ideas. The two sets of

scores are used together. There are a series of stages in the execution of a research design incorporating contextual/conceptual content analysis.

The first stage involves the methodological choice of content analysis. A contextual/conceptual content analysis is appropriate in one of three situations. Because it is basically an approach to measurement, it is useful in (a) descriptive or explanatory studies especially where one wants to identify and contrast meanings for one or more text units, (b) in hypothesis testing, or (c) in exploratory inquiries especially where questions are complex, uncharted and changing.

The second stage involves decisions on specific research procedures. Contextual/conceptual content analysis involves all the usual considerations in research such as design, measurement, sampling, pretesting, data-gathering, all of the possibilities of statistical analysis and reporting. In each of these, standard considerations about theoretical grounding and craftsmanship apply.

Since contextual content analysis examines patterns of use of ideas in text, it is important that the text qualify as research data. That is, it must be relevant to the research problem and contain characteristic patterns of word usage rather than an edited or altered pattern of usage. In the interview situation, skill in providing a free, natural stimulus to expression with minimal intrusive constraint is important. Use of a verbatim transcript (or a representative sample from it) is critical because it contains the pattern information central to contextual/conceptual analysis. A machine-readable computer file of the desired verbatim text is created. Word processors are useful for this purpose, and optical scanners are available which read printed text and convert it directly into a computer file.

The third stage involves the scoring procedures themselves. McTavish's model uses conceptual dictionary augmented with the four contexts (traditional, practical, emotional and analytic). The computer matches each word in the text against the word meanings in the dictionary, keeping a running tally of usage, concept by concept. Words not in the dictionary are tallied in a "leftover" list. Conceptual

category tallies are percentaged for each text by the total words in the text. This score is subtracted from an expected score obtained from a norm to yield an emphasis score for each of the concepts included in the dictionary. It is important to take account of variability in the use of ideas/words across social contexts. This is done by dividing by the standard deviation of category usage across the four contexts to yield useful emphasis scores (E-scores):

$$E\text{-score}(i,k) = (p_{i,k} - P_i) / S_i \dots\dots\dots(3.1)$$

where $E\text{-score}(i,k)$ is the E-score for category "i" in text "k"; $p(i,k)$ is the observed proportion of text in conceptual category "i" for text "k"; $P(i)$ is the overall expected probability of use of category "i"; and $S(i)$ is the expected standard deviation of category usage across contexts.

E-scores are computed for each of 116 idea/word categories distinguished in the current McTavish dictionary. They are the basic measures used for the *conceptual* analysis. The pattern of connectedness of various ideas in a text is examined using a clustering routine. Similarity and distinction between texts in terms of emphasized patterns of ideas can be quantified as well. A distance between texts can be measured as a discrepancy between texts on their profile of relative use of the 117 conceptual categories. The structure of conceptual differences shown in this proximity matrix can also be examined by clustering and other statistical techniques.

Four *C-scores* or contextual scores are also created during computer processing of the text. As each word is identified and classified into a conceptual category, four cumulative contextual scores are updated as illustrated in Table 2. The updating uses weights which reflect the relative use of each conceptual category in the four general social contexts. At any point during processing, these accumulating scores are available to be used in contextually disambiguating ambiguous words. Context scores are used to decide between alternative categorizations. Accumulated contextual scores over a text are standardized. These four scores are the four

contextual dimension measures. Distances between texts in this four-space can be computed and used to express the proximity of texts to each other, in terms of their approach to the ideas that are discussed. Cluster analysis helps display the structure of this proximity matrix.

Finally, E-scores often identify fruitful starting points for further qualitative analysis. The computer can further assist qualitative analysis by sorting and organizing the text, by searching for all instances of the use of some word or phrase, or by showing the use of key words in sentences and phrases in the text through KWIC lists. An inspection of these phrases often permits a refinement of the sense of the general conceptual categories and helps identify broader concepts extending across several conceptual categories. This grounding draws on strengths of qualitative approaches to text analysis within a systematic, comparative research framework.

Variables, including composite indices developed from contextual and conceptual analysis scores, can be included in a data set for statistical analysis, together with variables developed in any of the other more traditional ways.

Table 3.2 : Accumulation of C-Scores for the Illustrative Sentence

SENTENCE: "Work like mine keeps me from doing my best"

Accumulation of C-score Deviations

Words	Traditional	Practical	Emotional	Analytic	Average Weight
Work	0,001	0,008	-0,004	-0,003	0,009
like	0,001	0,006	-0,002	0,005	0,012
Mine	-0,008	0	0,018	-0,01	0,021
Keeps	-0,007	0,001	0,016	-0,01	0,027
Me	-0,007	-0,008	0,033	-0,018	0,038
From	-0,007	-0,002	0,016	-0,007	0,149
Doing	-0,008	0,002	0,016	-0,01	0,166
My	-0,008	-0,007	0,034	-0,019	0,177
Best	-0,007	-0,007	0,034	-0,02	0,184
C Score For Sentence	-7	-7	34	-20	0

In Table 3.2, words in text are looked up in a computerized dictionary and their idea category are identified. Probability of occurrence of that category in each "marker" context is added to accumulating context sums. In the illustration above deviations from the mean probability of the occurrence for that category are summed. Negative figures reflect below mean deviations. Positive figures indicate above mean deviations or an emphasis on that contextual approach. Final C- Scores for the entire sentence are multiplied by a constant, and their mean or a text is set at zero.

3.4 Procedures of Qualitative Content Analysis

3.4.1 Inductive category development

Classical quantitative content analysis has few answers to the question from where the categories come, how the system of categories is developed: "How categories are defined is an art. Little is written about it." [5]. But within the framework of qualitative approaches it would be of central interest, to develop the aspects of interpretation, the categories, as near as possible to the material, to formulate them in terms of the material. For that scope qualitative content analysis has developed procedures of inductive category development, which are oriented to the reductive processes formulated within the psychology of text processing.

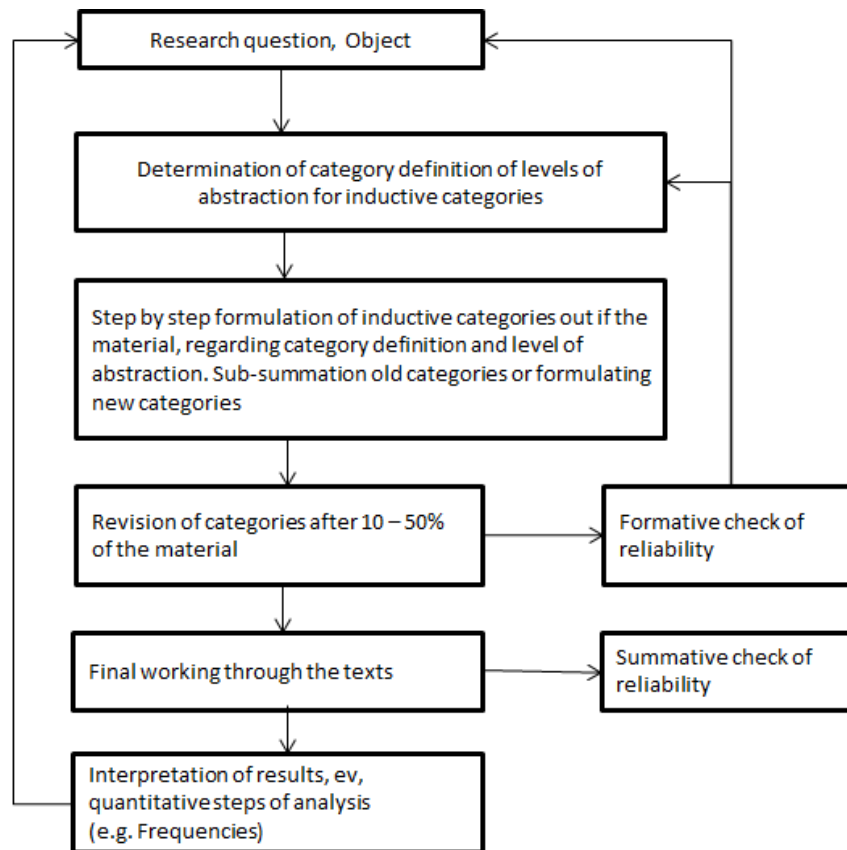


Figure 3.1: Step Model of Inductive Category Development

The main idea of the procedure is, to formulate a criterion of definition, derived from theoretical background and research question, which determines the aspects of the textual material taken into account. Within a feedback loop those categories are revised, eventually reduced to main categories and checked in respect to their reliability.

3.4.2 Deductive category development

Deductive category application works with prior formulated, theoretical derived aspects of analysis, bringing them in connection with the text. The qualitative step of analysis consists in a methodological controlled assignment of the category to a passage of text. Even if several procedures of text analysis are processing that step, it is poorly described. Here the step model within qualitative content analysis:

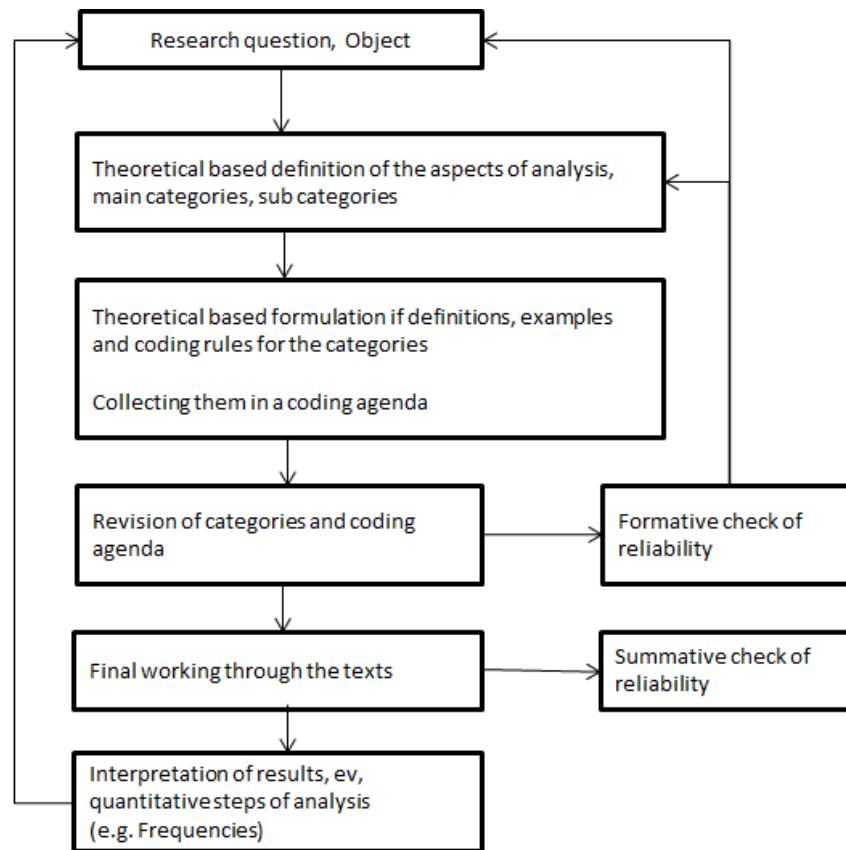


Figure 3.2: Step Model of Deductive Category Application

Then main idea here is to give explicit definitions, examples and coding rules for each deductive category, determining exactly under what circumstances a text passage can be coded with a category. Those category definitions are put together within a coding agenda. With the qualitative content analysis procedures of systematic text analysis are described, which try to preserve the strengths of content analysis in communication science (theory reference, step models, model of communication, category leaded, criteria of validity and reliability) to develop qualitative procedures (inductive category development, summarizing, context analysis, deductive category application) which are methodologically controlled. Those procedures allow a connection to quantitative steps of analysis if it seems meaningful for the analyst. The procedures of qualitative content analysis seem less appropriate,

- if the research question is highly open-ended, explorative, variable and working with categories would be a restriction, or
- if a more holistic, not step-by-step ongoing of analysis is planned.

As a result qualitative content analysis can be combined with other qualitative procedures.

CHAPTER 4

CATA – COMPUTER AIDED TEXT ANALYSIS

As it is mentioned before content analysis is to establish a common understanding of methodological assumptions. Content analysis is a summarizing, quantitative analysis of messages that relies on the scientific method, including attention to objectivity/intersubjectivity, a priori design, reliability, validity, generalizability, replicability, and hypothesis testing. It is not limited as to the type of messages that may be analyzed, nor as to the types of constructs that might be measured [10]. In this Chapter Computer Aided Text Analysis methods and software are discussed

4.1 Coding Methods in Text Analysis

There are two main options exist for the execution of quantitative content analysis. Human coding/Judge-based: The most commonly used type is that which uses human coders to analyze message characteristics. A coding scheme is modeled by researchers, and the instrument is applied to message content by trained, reliable coders.

Computer- Aided Text Analysis (CATA): Using computer applications to analyze text, introduces a growing set of options for automated analyses.

However, human coding procedures are important for:

- The origination of content analytic schemes that eventually become CATA algorithms
- The measurement of highly latent constructs
- The ongoing validation of CATA measures

Reliability of CATA is in no way comparable to human coding. Of course, an extra period for machine learning is necessary for any CATA system. But this becomes insignificant compared to the reliability provided by CATA.

4.2 CATA Tools

CATA began with the General Inquirer, a mainframe computer application introduced by Philip Stone of Harvard in 1965. The purpose of the General Inquirer was to automate the analysis of textual information, searching for text that delineates such features as valence, Osgood's three semantic dimensions and language reflecting particular institutions, emotion-laden words, cognitive orientation, and more. All CATA programs have their basis as the analysis of text via the application of some algorithms of word or word sequence searching and counting. Most often, the analysis involves one or more dictionaries, i.e., lists of search terms intended to measure constructs on the text.

Two prominent CATA programs that include well-documented pre-set dictionaries are LIWC and Diction 5.0. In LIWC 2007 (Linguistic Inquiry and Word Count) there are 84 dictionaries that tap such linguistic and semantic concepts as use of first-person pronouns, anger, optimism, reference to home, and reference to motion. The program Diction 5.0 (Hart, 2000), designed to analyze political speech, has 31 pre-set dictionaries, including those intended to measure tenacity, aggression, praise, satisfaction, and complexity. The 31 dictionaries are also combined to form "master variable" scales: Activity, optimism, certainty, realism, and commonality. The alternative to using pre-set dictionaries is to create one's own custom dictionaries, and most CATA programs allow for this. However, the development of original dictionaries is quite demanding and ought to include a validation process that links measured dictionaries with additional indicators of the construct under investigation.

4.3 Quantitative and Qualitative CATA Tools

Generally Quantitative CATA tools reads text and produces a variety of outputs ranging from simple diagnostics such as word and alphabetical frequencies to a summary of the "main ideas" in a text. It uncovers patterns of word usage and produces such outputs as simple word counts, cluster analysis, and interactive some neural analysis. Some of them creates 2D or 3D concept maps for further analysis. Some CATA tools have the Qualitative Analysis functionality to make further analysis to classify text for emotions such as Activity, Optimism, Certainty, Realism and Commonality. These tools generally used for political or psychological analysis. A list is given in Table 4 regarding purpose of the programs.

Table 4.1 : CATA Programs

Program	Author	Original Purpose	Type
VBPro	M. Mark Miller	Newspaper articles	Word count/researcher-created dictionaries only
Yoshikoder	Will Lowe	Political documents	Word count/researcher-created dictionaries only
WordStat	Normand Peladeau	Statistical analysis package	Word count/researcher-created dictionaries only
General Inquirer	Philip Stone	Mainframe computer application	Word count with pre-set dictionaries
Profiler Plus	Michael Young	Communications of world leaders	Word count with pre-set dictionaries
LIWC 2007	Pennebaker, Booth, & Francis	Linguistic characteristics & psychometrics	Word count with pre-set dictionaries (researcher-created dictionaries may be added)
Diction 5.0	Rod Hart	Political speech	Word count with pre-set dictionaries
PCAD 2000	Gottschalk & Bechtel	Psychiatric diagnoses	Word count with pre-set dictionaries (researcher-created dictionaries may be added)
WORDLINK	James Danowski	Network analysis/communication	Word co-occurrence
CATPAC	Joseph Woelfel	Consumer behavior/marketing	Word co-occurrence

CHAPTER 5

PROPOSED SOLUTION FOR ANALYSIS OF TURKISH NEWS WEBSITES

As it is mentioned before, social and economic developments needs to be analyzed carefully and quickly. Especially companies, politicians try to learn what is going on in the suburban area. PR companies in last few years try to give some localized news summaries but they are lack of website analysis. On the other hand Turkish Language has some difficulties to analyze in terms of content, because of Turkish Characters. These characters also cannot be interpreted in current software. Also there are some difficulties of keeping the standards in web pages. So a solution that enables analysis of local news websites with accuracy and speed is a need. Also software that analyzes the website with Turkish Character set should be considered. In this chapter building blocks and concept design will be given.

5.1 Challenges

Currently there are number of content analysis tools in the market or academic area. The most important thing is to first standardize the website according to the rules that will be run in the software. These tools or worldwide search engines has some disadvantages like not being local. In this project, the developed software is designed to overcome the following challenges those are faced in current solutions:

- Focus: In a search engine or content analysis software you cannot define a search area like Turkey's Local News Websites
- Speed: Current search engines can create indexes within 24 hours which is not satisfactory
- Language: Local (Turkish) Character Set support is important
- Objectivity: Current search engines or solutions change their analysis commercially.

- Accuracy: Advertisements, comments can disturb search result.
- Local Information: No analysis tool has local geographic information.

5.2 Technologies Used

In the current project because of flexibility and easiness of development Microsoft Technologies are used. Technologies and tools used are:

- Microsoft SQL Server 2008 for the database
- Microsoft Full Text Search Engine for search
- .Net Framework 3.5 is used for the runtime environment
- Visual Studio 2010 is used for development
- C# as programming language

In terms of design patterns two design patterns are preferred. Singleton Pattern is preferred for the database operations and crawler job. Purpose of the usage of this pattern is there are times when an application needs a single instance of a given class and a global point of access to that class. Because especially services and database connections needs global management Singleton Design pattern is used. On the other hand Factory Method is used in Page Fetcher and Analyzer because of the need of creating many object instances in the analysis.

5.3 Conceptual Design

In the developed solution main building blocks are given in Figure 4. The building blocks of the concept are:

- Crawler
- Text Search
- Page Analyzer
- Database
- Services

5.3.1 Crawler

Crawler is mainly responsible for collecting data from the Internet, The list of the news websites comes from the database which can be updated from the news associations. First page analysis is done in this component. Also some alphabetic or character set problems are resolved here.

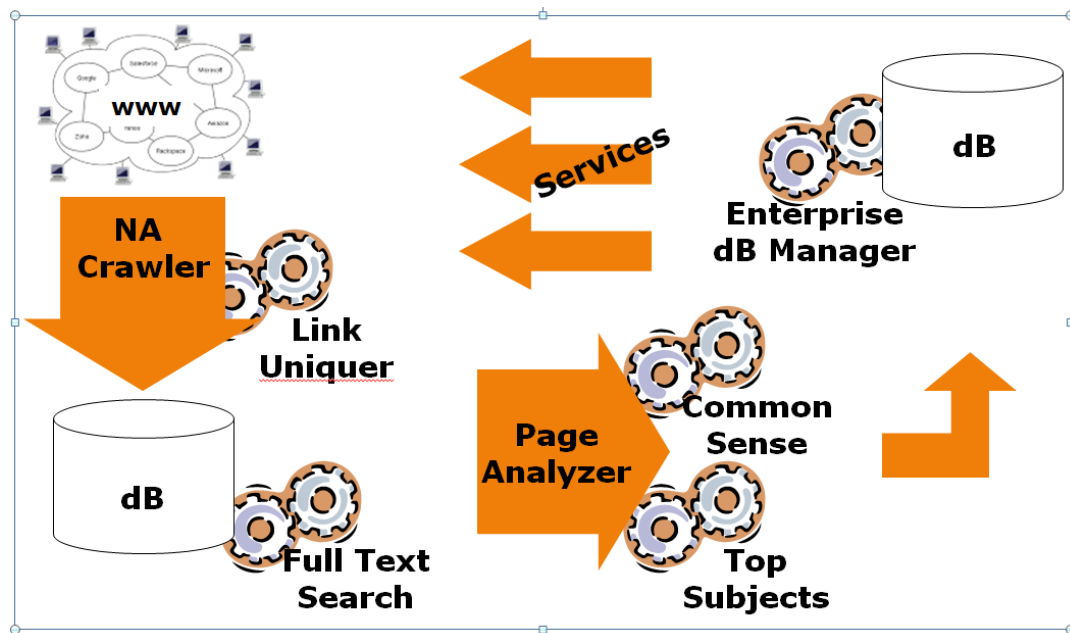


Figure 5.1: Conceptual Design for the Solution

5.3.2 Text search

Text search is the engine that enables the search in the collected and clarified content. In this project standard full text engine of Microsoft is used.

5.3.3 Page analyzer

Page Analyzer analyzes html code and makes the further operations such as the website is returning an error, is there any advertisements in the website, are the links external or internal. Data is inserted in the database by the crawler according to the website geographical information. Also Alexa and Google Page rank from

the internet is taken and localized statistics are created according to the category, popularity and geographic distribution.

5.3.4 Database

All the content collected, error logs, page rankings exist in the database. Also geographical distribution of the websites exists in the database.

5.3.5 Services

Services are mainly responsible for running operations like scanning websites, listening other collection crawler jobs and getting information from Alexa website, link unification.

CHAPTER 6

EXPERIMENTAL STUDIES AND RESULTS

In this chapter the results of the experimental studies are presented. In this study because the main focus is overcoming the problems in nonstandard websites, page analysis algorithm and code is given.

6.1 Crawler

Crawler is the first step to analyze to clear the content. In this step web page is being analyzed, if there are web site errors they are discarded, advertisements are cleared. Additionally, Turkish character problems are solved.

Crawler (mentioned as NaCrawler in the study) is the program that takes the html code of the web page, that is listed in the database. First of all, html code is taken and saved to database. After that application starts searching internal and external links. Internal links are queued to be analyzed in the next stage. External links are saved to the database for further analysis. If the web page is returning an error regarding that page cannot be displayed, url is not found etc, this error is recorded. In the appendix developed code is given. The basic flowchart is given in Figure 4 and steps of the application are as follows:

- Step1: It is checked whether another crawler is working in the page
- Step2: First Page is analyzed and all the links are taken out (AnalzStartPage). Links those cannot return text are defined as exceptions (i.e. wav, mp3).

- Step3: If the page analysis returns an error, for further debugging link of the page is saved with the specific error. If the next checking time is less than 3 hours, this page is discarded for analysis.
- Step4: Link Uniquer controls whether the links will be checked or not.
- Step5: If there are less than 10 links in the page form, the database previous searches completes the number of samples to 10.
- Step6: Sample pages are compared with each other and the content is unified. Pages are cleaned from advertisements with a logic of comparing last 10 samples. If the same text is investigated in 3 pages, this content is decided as advertisement. Contents are checked according to standard character sets. If some non-Latin Character exists in the content (i.e. a??lama instead of açıklama) it is corrected with Turkish Characters.
- Step7: If the link returns an error, it is discarded
- Step8: Page information is saved to the database

6.2 Infrastructure Information

Application developed for web page analysis is mainly based definitions that are given in database . In the designed application, mainly there are eight definition groups. These groups mainly contain Infrastructure information for the application to run. Also some statistical data exists in these schemas which are given with sample data as follows:

- **Website List:** In this list web sites that are planned to be analyzed are given. As shown in Table 5, websites are classified according to their Names, Url Address, City Plate Number, Telephone Numbers and Contact E-mail.
- **Webpage List:** Web pages under web sites are classified according to their content such as Latest News, Sports, Politics Economy, Environment etc. In Table 6.2, a sample webpage list under a website is given. Also last analysis time is logged in this table.

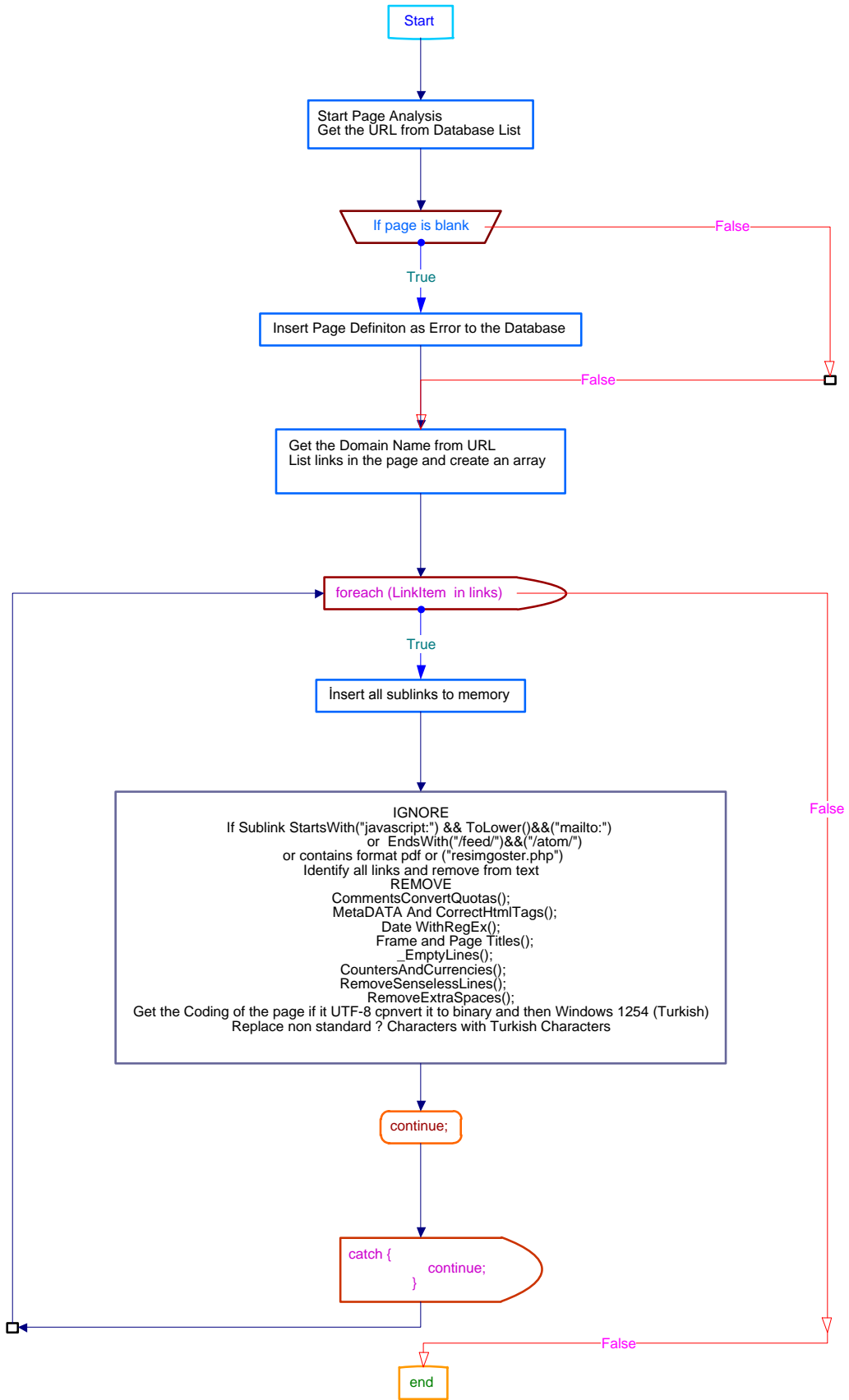


Figure 6.1: Flowchart of the Page Analysis Class

Table 6.1: Web Site List Schema

OID	Name	Cit y ID	Sub Group ID	Content Area ID	Description	Phone	Phone 2	Adress	Email
2271	Adana 5 Ocak	1	3	4	http://5ocak.com.tr/tr/	NULL	NULL	NULL	NULL
2272	Adana Aktüel	1	3	4	http://www.adanaaktuel.com.tr/	NULL	NULL	NULL	NULL
2273	Adana Haber	1	2	4	http://www.adanahaber.com.tr/				

Table 6.2: Web Page List and Categories

OID	Name	Main Definition ID	Content Type ID	Link	LastFetch Time	Priorit y	Statu s ID	Fet ch Ty pe ID
1	GÜNCEL	2271	1	http://5ocak.com.tr/tr/k.asp?kid=174	31.10.2011 07:22	5	1	0
2	ASAYIŞ	2271	1	http://5ocak.com.tr/tr/k.asp?kid=309	31.10.2011 05:45	5	1	0
7	Anasayfa	2271	1	http://5ocak.com.tr/tr/	31.10.2011 05:17	5	1	0
8	Özel Haber	2271	1	http://5ocak.com.tr/tr/k.asp?kid=290	31.10.2011 07:22	5	1	0
9	Siyaset	2271	1	http://5ocak.com.tr/tr/k.asp?kid=280	31.10.2011 07:22	5	1	0
10	Ekonomi	2271	2	http://5ocak.com.tr/tr/k.asp?kid=279	31.10.2011 07:22	5	1	0
11	Son Dakika	2271	2	http://5ocak.com.tr/tr/k.asp?kid=174	31.10.2011 07:22	5	1	0
12	Çevre	2271	1	http://5ocak.com.tr/tr/k.asp?kid=281	31.10.2011 07:22	5	1	0
14	Spor	2271	5	http://5ocak.com.tr/tr/k.asp?kid=218	31.10.2011 07:22	5	1	0
15	Kültür Sanat	2271	8	http://5ocak.com.tr/tr/k.asp?kid=219	31.10.2011 05:05	5	1	0
16	Yaşam	2271	1	http://5ocak.com.tr/tr/k.asp?kid=234	31.10.2011 05:05	5	1	0
17	Asayiş	2271	1	http://5ocak.com.tr/tr/k.asp?kid=309	31.10.2011 07:22	5	1	0
18	Rehber	2271	10	http://5ocak.com.tr/tr/k.asp?kid=258	31.10.2011 07:22	5	1	0

- **Content Area:** In this schema content areas such as local, national or international are given.
- **Content Type:** This Schema gives main content categories those are used in web page content type classification. In Table 6.3, list for content type is given.

Table 6.3: Content Type

OID	Name
1	Haber
2	Ekonomi
3	Eđitim
4	Magazin
5	Spor
6	Turizm
7	E-Ticaret
8	Kltr ve Sanat
10	Diđer
11	Bilim ve Teknoloji
12	Sađlık
13	Otomotiv

- **Geographical Areas:** Seven Geographical Areas of Turkey are defined here. Also Cyprus Turkish Republic is also defined. This information is used for localized analysis according to the geographic areas.
- **Cities:** City list is given with Traffic Code, Geographical and Statistical Area Code, Population and Total Income according to the National Statistics Institute data. In table 6.4, sample for this list is given.

Table 6.4: City List

OID	Country ID	Name	Traffic Code	Geographical Area	Statistical Area	Population	Acreage	GSYH Rate
1	1	Adana	1	3	6	2711368	14256	3
2	1	Adiyaman	2	7	10	623811	7572	0,4
3	1	Afyon	3	2	4	812416	14532	0,7
4	1	Ađrı	4	6	12	528744	11315	0,2
5	1	Amasya	5	5	7	365231	5731	0,4
6	1	Ankara	6	4	2	4007860	25615	7,6
7	1	Antalya	7	3	6	1719751	20599	2,6

- **Site Groups and Site Subgroups:** In this schema list contains main classes of sites according to their type such as Internet Media, Naitonal Web Media,

News Agency, Blog, Social Network, Forums etc. . Table 6.5, has full details of the grouping.

Table 6.5: Site Group List

OID	SiteGroupID	Name
1	1	Yaygın Medya Siteleri
2	1	İnternet Medyası
3	1	Sektörel Medya
4	1	Haber Ajansları
5	2	Kişisel Blog
6	2	Kurumsal Blog
7	2	Topluluk Siteleri
8	2	Forum
9	3	Kurumsal Siteler
10	3	Ticari Siteler
11	3	Diğer

- **Education Status:** Sample from Educational statistics of all cities are given in table 6.6. This data is planned to be used in social analysis of the audience.

Table 6.6: Educational Statistics

CityID	Okuma Bilmeyen	Okula Gitmemis	İlkokul	İlkOgretim	Lise	Univ	Lisans Doktora
1	8,4525	22,6395	27,3267	16,8177	17,7651	6,5657	0,433
2	13,1866	27,4797	23,1185	18,9877	13,2797	3,8019	0,1459
3	7,559	20,1275	36,7911	16,607	13,5902	4,9689	0,3563
4	16,1867	46,6587	15,6461	11,8773	7,4603	2,0199	0,1509
5	6,8257	18,7286	35,3315	16,6719	15,7239	6,4736	0,2448
6	3,6795	16,8967	24,8664	16,5334	23,279	13,0102	1,7349
7	4,9166	18,3706	33,7581	16,3754	17,9856	8,1603	0,4334
8	7,0166	17,9127	33,4713	16,1629	19,1268	6,0482	0,2614
9	7,5785	18,2651	37,4695	14,8235	14,5967	6,9279	0,3388
10	6,4466	16,6337	38,3122	14,9863	16,0735	7,1598	0,3879
11	3,9342	15,8319	37,9341	15,8942	19,9001	6,228	0,2775

6.3 Content Deduction

In this study one of the most important functions is cleaning the content and getting ready for the analysis. Although a lot of algorithms are used in the application, basically there are two main content deduction and two main correction methods.

These are:

- **Correction of character set problems:** Html code is checked and if it starts with UTF format, all characters are converted to binary and then Turkish character set.
- **Correction of common misusages:** A correction list can be created in the database for common misusages like istanbul instead of İstanbul. This list is created and managed manually.
- **Regex Removal:** Regex class is responsible of the removal of data which is not related directly with the main content such as counter information, page numbers, currency, date and online visitor statistics. A list of removed content is given in Table 6.7.

Table 6.7: Regex Removal List

OID	RegExQuery
1	((bugün [0-9]{1,7}) {1,3}(dün [0-9]{1,7}) {1,3}(bu hafta [0-9]{1,7}))
2	(\\$ {0,2}\d{0,1}(\.\d{1,5})?) {1,3}(€ {0,2}\d{0,1}(\.\d{1,5})?)
3	(dolar \d{0,1}(\.\d{1,5})?) {1,3}(euro \d{0,1}(\.\d{1,5})?)
4	(\\$ {0,2}\d{0,1}(\.\d{1,5})?) {1,3}(€ {0,2}\d{0,1}(\.\d{1,5})?)
5	(dolar \d{0,1}(\.\d{1,5})?) {1,3}(euro \d{0,1}(\.\d{1,5})?)
6	((ziyaretci) {1,3}(bugün [0-9]{1,7}) (toplam [0-9]{1,7}))
7	((bugün : [0-9]{1,7}) {1,5}(toplam : [0-9]{1,7}))
8	((bugün {1,3}[0-9]{1,7}) {1,5}(toplam {1,3}[0-9]{1,7}))
9	sitemizi bugüne kadar toplamda [0-9]{1,10} kişi ziyaret etmiştir.
10	(bugün çoğul [0-9]{1,10} {0,4}toplam çoğul [0-9]{1,10})
11	(şu anda online {0,4}[0-9]{1,5} {0,2}kişi {0,3}bugünkü ziyaret {0,3}[0-9]{1,5} {0,2}kişi {0,3}toplam ziyaret {0,3}[0-9]{1,10} {0,2}kişi)
12	(online ziyaretçi : {0,4}[0-9]{1,3} {0,3}toplam : [0-9]{1,10})
13	(şu an sitede [0-9]{1,10} kişi var)
14	(alış satış) {1,3}: {1,3}[0-9].[0-9]{2,4}
15	euro {2,5}dolar

- **Html Code Clean:** Html page has some standard notations and to get the real content from the html code, coding notations are removed from the content. A list of removed content is given in Table 6.8.

Table 6.8: Code Removal List

OID	StartsWith	EndsWith	Açıklama
1	<style	/style>	objenin style'ını belirler, font, büyüklük , renk vs.
2	<marquee	/marquee>	Kayan nesneyi hareket ettirir(kayan yazı, fotoğraf vs)
3	<marquee	/marquee>	Kayan nesneyi hareket ettirir(kayan yazı, fotoğraf vs)
4	<map	>	İmaj üzerinde belirli kordinatlar üzerine işlemler yapar.
5	<select	</select>	Dropdown nesnesini oluşturur (Combobox)
6	<option	</option>	Dropdown nesnesinin öğeleri yazılır.
7	<script	</script>	Html içerisinde script yazma yeridir
8	<script	</script>	Html içerisinde script yazma yeridir
9	<noscript	</noscript>	Script desteklemeyen browserlar için kullanılır
10	<noscript	</noscript>	Script desteklemeyen browserlar için kullanılır
11	<input	>	Kullanıcıdan değer alır.
12	<button	>	Buton oluşturur
13	<link	>	Doküman ile harici kaynak arasında ilişki oluşturur
14	<link	>	Doküman ile harici kaynak arasında ilişki oluşturur
15	<embed	</embed>	Sayfaya browser plugin'i koymaya yarar.
16	<noembed	</noembed>	embed objesini deteklemeyen browserlar için kullanılır.
17	<object	</object>	Activex Nesneleri sayfaya gömmek için kullanılır. Örneğin flash player bununla eklenir sayfaya)
18	<iframe	</iframe>	Mevcut sayfa içinde başka bir sayfayı göstermek için kullanılır)
19	<textarea	>	Metin gösterir.
20	<frame	>	birbirinden bağımsız sayfaları, aynı sayfa içerisinde göstermek için kullanılır.
21	<area	>	map komutundaki gibi imajı kordinatlandırır.
22		görsel nesnesi eklenir. (Fotoğraf)
23	<ul		madde imleme
24	<div	>	Kendine özgü özellikleri olan bölüm oluşturur
25	<div	>	div ile aynı
26	<table	>	Tablo oluşturur.
27	<tbody	>	Tablo başlığıdır
28	<td	>	Tablo sütunu
29	<tr	>	Tablo satırıdır
30	<form	>	Dataları sunucuya göndermek için kullanılır
31	<h1	>	Başlık 1 Nitelendirmesi
32	<h2	>	Başlık 2
33	<h3	>	Başlık 3
34	<h4	>	Başlık 4

35	<h5	>	Başlık 5
36	<h6	>	Başlık 6
37	<hr	>	Altçizgi ekler
38		bullet oluşturur. (List)
39		Tek satır yazı
40	<html	>	Html başlangıç işareti
41	<body	>	Body başlangıcı Sayfa içeriği burada yer alır.
42	<head	>	html'in Scriptler meta ve referans linkleri bu kısma yazılır
43	<dl	</dl>	dlt ve dt tagları ile definition List oluşturur. (tablı şekilde yazılmış yazılar)
44		Kalın yazı yazdırır
45	<th	>	Tablo Başlığı
46		Font belirtir.
47	<?xml	</xml>	xml dosyanın başlangıç ve bitişini belirtir.
48	<ilayer	>	Yazı'nın ilgili satırda pozisyonunu belirtir.
49	<layer	>	Elementlerin pozisyonlarını ve hareketlerini ayarlar.
50	<base	>	Sayfa için varsayılan url ve hedef belirtir.
53	<label	>	Giriş elementi için etiket belirtir.
54	<small	>	Küçük yazı yazdırır.
55	<i	>	İtalik yazı yazdırır.
56		Kalın yazı yazdırır.
57	<embed	>	Browser plugin'i koyar. (Client tarafındaki uygulama)
58	<col	>	Tablodaki kolon ya da kolonların özelliklerini tanımlar.
59	<bgsound	>	Sayfa açıldığında müzik çalar.
60	<script	</ script>	Bu kısma script yazılır (Javascript , vbscript)
61	class	>	Nesne'nin class'ını belirtir. O class'tan özellikleri devralır.
62	<eval	>	İlgili öğeye değer vermek için kullanılır.
63	<a	>	Link Vermek için kullanılır.

6.4 Results

After correction and removal procedures content becomes ready for search. For example as given in Figure 5 <http://www.malatyahavadis.com/h9805-prestiji-kalsin-once-bitirin.html> web page is taken and saved in the database (Figure 6). After correction and removal of the html headers and miscoding content becomes suitable for indexing. In Table 13, unprocessed and processed content is given. As it is seen from the output, all data other than the main content is removed. This algorithm is currently being used in www.netajan.com news analysis website .

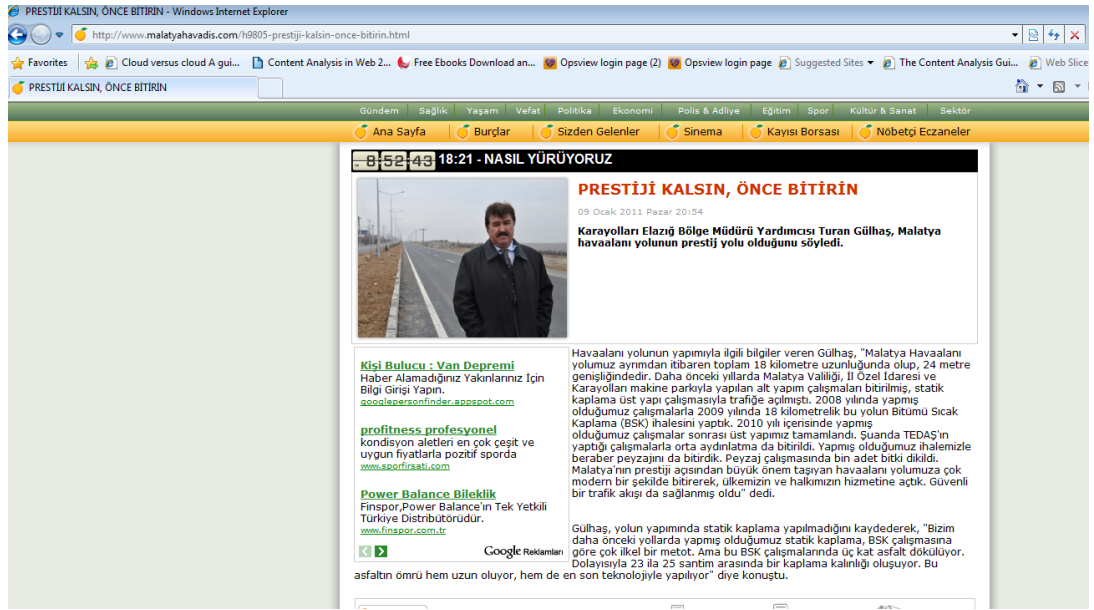


Figure 6.2 : Sample Web Page

OID	PageDefinitionID	Title	Link	SaveDate	SaveTime	Searched	ContentAreaID
185	4479	habere cevap niye adresinden gelmedi? referandum çmaları	http://www.habertrakya.com/index.php?option=com_content&view=ar	10.01.2011	11:30:40	0	4
186	8521	kurucular kurulu üyeleri s.no adı soyadı s.no adı soyadı a g	http://www.mardav.org.tr/?lg=fe2705e873c5cc6ea04b74c7009c4046219f6c	10.01.2011	11:30:41	0	4
187	4479	aynca maxi veriş merkezinin 3 katında yapılan	http://www.habertrakya.com/index.php?option=com_content&view=ar	10.01.2011	11:30:41	0	4
188	8521	m.nihat iskenderoğlu 2001 ve 2004 selahaddin aydođu 200	http://www.mardav.org.tr/?lg=fe2705e873c5cc6ea04b74c7009c4046219f6c	10.01.2011	11:30:41	0	4
189	4479	referandum çmaları kapsamında 23 ağustos pazartesi günü	http://www.habertrakya.com/index.php?option=com_content&view=ar	10.01.2011	11:30:41	0	4
190	4479	belediye başkanı n bilmen kaçınıcı kez tanıtım	http://www.habertrakya.com/index.php?option=com_content&view=ar	10.01.2011	11:30:42	0	4
191	8521	normal üyeler üyeler a.halim günüşü a.halim ölmen a.halim	http://www.mardav.org.tr/?lg=fe2705e873c5cc6ea04b74c7009c4046219f6c	10.01.2011	11:30:42	0	4
192	4923	çlip karşivör 09 ocak 2011 pazar 21:02 cumhuriyet halk	http://www.malatyahavadis.com/h9809-clip-karisivor.html	10.01.2011	11:30:42	0	4
193	4923	ak parti yöneticileri, köy ziyaretlerinde 30 aralık 2010	http://www.malatyahavadis.com/h9702-ak-parti-yoneticileri-koy-ziyaretleri	10.01.2011	11:30:43	0	4
194	8521	e-bilgi edirne formu applicanon form tüzel kişinin	http://www.mardav.org.tr/?lg=fe2705e873c5cc6ea04b74c7009c4046219f6c	10.01.2011	11:30:43	0	4
195	4479	belediye adından başka yerde olmayan başından sonuna	http://www.habertrakya.com/index.php?option=com_content&view=ar	10.01.2011	11:30:43	0	4
196	4912	belediye otobüsleri zamlandı 09 ocak 2011 pazar 20:59 mala	http://www.malatyahavadis.com/h9808-belediye-otobusleri-zamlandi.html	10.01.2011	11:30:43	0	4
197	4923	malatya büyükşehir olmaya hak kazandı 27 aralık 2010	http://www.malatyahavadis.com/h9677-malatya-buyuksehir-olmaya-hak-ka	10.01.2011	11:30:43	0	4
198	4479	yaman adam dediğ ya, gerçekten de öyle. ilçe	http://www.habertrakya.com/index.php?option=com_content&view=ar	10.01.2011	11:30:43	0	4
199	8521	burs hakkında burs alanların dikkatine : vakfa uğrayın	http://www.mardav.org.tr/?lg=fe2705e873c5cc6ea04b74c7009c4046219f6c	10.01.2011	11:30:43	0	4
200	4912	fiat bey diyorki 09 ocak 2011 pazar 20:58 malatya'nın	http://www.malatyahavadis.com/h9807-fiat-bey-diyorki.html	10.01.2011	11:30:43	0	4
201	8521	ücretsiz hukuki danışmanlık hizmeti mardav-adana'daki mar	http://www.mardav.org.tr/?lg=fe2705e873c5cc6ea04b74c7009c4046219f6c	10.01.2011	11:30:44	0	4
202	4912	ak partililer pütürgeçiler derneğinde buluştu 09 ocak 2011	http://www.malatyahavadis.com/h9806-ak-partililer-puturgeciler-derneğinde-	10.01.2011	11:30:44	0	4
203	4923	siyasetle ilgili kendime bir hedef koymadım 25 aralık	http://www.malatyahavadis.com/h9666-siyasetle-ilgili-kendime-bir-hedef-ko	10.01.2011	11:30:44	0	4
204	4479	8 den ayrılak neredeyse 1 sene olacak.	http://www.habertrakya.com/index.php?option=com_content&view=ar	10.01.2011	11:30:44	0	4
205	4479	sayfa 1 topal ördek (lame duck	http://www.habertrakya.com/index.php?option=com_content&view=ar	10.01.2011	11:30:45	0	4
206	4912	prestiji kalsın, önce bitirin 09 ocak 2011 pazar	http://www.malatyahavadis.com/h9805-prestiji-kalsin-once-bitirin.html	10.01.2011	11:30:45	0	4

Figure 6.3: Content Title and Information in the Database

Table 6.9: Processed and Unprocessed Content

OID	PageContent	PageSource
10726 48	<p>havaalanı yolunun yapımıyla ilgili bilgiler veren gülhaş malatya havaalanı yolumuz ayırımdan itibaren toplam 18 kilometre uzunluğunda olup 24 metre genişliğindedir daha önceki yıllarda malatya valiliği il özel idaresi ve karayolları makine parkıyla yapılan alt yapım çalışmaları bitirilmiş statik kaplama üst yapı çalışmasıyla trafiğe açılmıştı 2008 yılında yapmış olduğumuz çalışmalarla 2009 yılında 18 kilometrelik bu yolun bitümü sıcak kaplama (bsk) ihalesini yaptık 2010 yılı içerisinde yapmış olduğumuz çalışmalar sonrası üst yapımız tamamlandı şuanda tedarik yaptığı çalışmalarla orta aydınlatma da bitirildi yapmış olduğumuz ihalemizle beraber peyzajını da bitirdik peyzaj çalışmasında bin adet bitki dikildi malatyanın prestiji açısından büyük önem taşıyan havaalanı yolumuza çok modern bir şekilde bitirerek ülkemizin ve halkımızın hizmetine açtık güvenli bir trafik akışı da sağlanmış oldu dedi gülhaş yolun yapımında statik kaplama yapılmadığını kaydederek bizim daha önceki yollarda yapmış olduğumuz statik kaplama bsk çalışmasına göre çok ilkel bir metot ama bu bsk çalışmalarında üç kat asfalt dökülüyor dolayısıyla 23 ila 25 santim arasında bir kaplama kalınlığı oluşuyor bu asfaltın ömrü hem uzun oluyor hem de en son teknolojiyle yapılıyor diye konuştu</p>	<pre> html> <head> <title>PRESTİJİ KALSIN, ÖNCE BİTİRİN</title> <meta http-equiv="Content-Type" content="text/html; charset=iso-8859-9" /> <meta name="keywords" content="Karayolları Elazığ Bölge Müdürü Yardımcısı Turan Gülhaş, Malatya havaalanı yolunun prestij yolu olduğunu söyledi. " /> <meta name="description" content="PRESTİJİ KALSIN, ÖNCE BİTİRİN" /> <meta name="verify-v1" content="PQtBgxandjgncQ11p3ziU/l4TP9vSA3/tX21rXDDp E=" /> <link href="ms_styles.css" rel="stylesheet" type="text/css" /> <link rel="shortcut icon" href="favicon.ico" /> <script type="text/javascript">function handleError() { return true; } window.onerror = handleError;</script> <script type="text/javascript" src="altmenu.js"></script> <script type="text/javascript" src="flash.js"></script> </head> <body bgcolor="#E6EAE0"> <table width="100%" cellpadding="0" cellspacing="0"> <tr> <td width="25%"></td> <td width="50%"> <table width="800" align="center" cellpadding="0" cellspacing="0"> <tr> <td colspan="3" height="25"></td> </tr> <tr> <td></td> <td width="750" background="images/ust_arka.jpg"></td> <td></td> </tr> <tr> <td background="images/sol_arka.jpg"></td> <td> <table width="100%" cellpadding="0" cellspacing="0" bgcolor="#FFFFFF"> <tr> <td width="216"> </td> <td> <table width="100%" cellpadding="0" cellspacing="0"> <tr> <td></td> <td></td> <tr> <td align="center"> <table width="468" height="60" class="cerceve"> <tr> <td><script type="text/javascript">strFlash(468, 60, "gozde.swf")</script></td> </tr> </table> </td> <td> <table> <tr> <td> <td> <table> <tr> <td background="images/sag_arka.jpg"></td> </tr> </table> </td> <td width="25%"></td> </tr> <tr> <td background="images/orta_dis_arka.jpg" width="25%"></td> <td> <table width="800" cellpadding="0" cellspacing="0" align="center"> <tr> <td></td> <td width="750" background="images/orta_arka.jpg" valign="top"> </pre>

CHAPTER 7

CONCLUSION

Content Analysis is a hot topic for politicians, companies, social analysts, doctors, psychiatrists and other social engineers that need to analyze people. Especially latest social movements in Middle East Region shows us that Internet Media is strongly used to manage crowd. On the other hand, local news websites generally hard to find because of their lowest rank in popularity index. Internet search engines are currently becoming commercial media for advertising and PR activities. Because this issue local web sites are in lower places in the ranking and being underestimated then their real effect on the people. Also local news websites has some common problems like lack of quality and standard in coding which makes difficult to analyze these web pages. As a summary challenges in the project are

- Focus: In a search engine or content analysis software you cannot define a search area like Turkey's Local News Websites
- Speed: Current search engines can create indexes within 24 hours which is not satisfactory
- Language: Turkish Character Set support is important
- Objectivity: Current search engines or solutions change their analysis commercially.
- Accuracy: Advertisements, comments can disturb search result.
- Local Information: No analysis tool has local geographic information.

These challenges have been successfully achieved and realized in this project. Content has become ready for search with geographic and demographic information of Turkey.

7.1 Future Work

An area of future development may be the enhancement of analysis tools and enriching the analysis with rankings to figure out likes and dislikes. Current database design is ready for ranking points. As a future work Alexa and Google ranks combined with some criteria like having a high number of original news, having a short and understandable URL name, number of copied news will enrich the like or dislike rating. On the other hand for Qualitative Analysis to understand feelings and attitudes in the news, a Turkish Dictionary can be created in order to analyze feelings or attitudes in the content. Dictionary can be classified according to emotional categories. Combining the dictionary with C-Score and E- Score more accurate and fast qualitative content analysis can be achieved.

REFERENCES

- [1] **Barney G. Glaser, Anselm L. Strauss** (1999) *The Discovery of Grounded Theory: Strategies for Qualitative Research* ISBN: 0202302601

- [2] **Professor Ian Budge, Hans-Dieter Klingemann, Andrea Volkens and Judith Bara** (2001) *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945-1998* Oxford University Press

- [3] http://en.wikipedia.org/wiki/John_Naisbitt (10.2011)

- [4] **Holsti, O.R.** (1969). *Content Analysis for the Social Sciences and Humanities*. Reading, MA: Addison-Wesley.

- [5] **Krippendorff, K.** (1980). *Content analysis: An Introduction to Its Methodology*. Beverly Hills, CA: Sage

- [6] http://en.wikipedia.org/wiki/Zipf's_law (10.2011)

- [7] **Weber, Robert Philip.**(1990) *Basic Content Analysis*, Second Edition. Newbury Park, CA: Sage Publications

- [8] **Krippendorff,**(2004) *Content Analysis, An Introduction to Its Methodology* 2nd Edition; Thousand Oaks, CA: Sage Publications

- [9] **McKeone, Dermot** (1995) *Measuring Your Media Profile*. Gower Publishing

- [10] **Neuendorf, K. A., Atkin, D., & Jeffres, L. W.** (2002). *Adoption of Audio Information Services in the United States*. Hampton Press, Inc.

- [11] <http://academic.csuohio.edu/kneuendorf/content/> (10.2011)
- [12] http://en.wikipedia.org/wiki/Content_analysis (10.2011)
- [13] **Berelson, B.** (1952). *Content Analysis in Communication Research*. New York: Free Press.
- [14] **Berelson, B., & Lazarsfeld, P.** (1948). *The Analysis of Communication Content*. Chicago and New York: University of Chicago and Columbia University.
- [15] **Gerd G. Kopper, Albrecht Kolthoff, and Andrea Czepek,** (2000) “*Research Review: Online Journalism—A Report on Current and Continuing Research and Major Questions in the International Discussion,*” *Journalism Studies*
- [16] **Guido H. Stempel III and Robert K. Stewart,** (2000) “*The Internet Provides both Opportunities and Challenges for Mass Communication researchers,*” *Journalism & Mass Communication Quarterly*
- [17] **Christopher Weare and Wan-Ying Lin** (2000; 18) *Content Analysis of the World Wide Web: Opportunities and Challenges Social Science Computer Review*
- [18] **Sally J. McMillan,** (Spr 2000) “*The Challenge of Applying Content Analysis to the World Wide Web.*” *Journalism and Mass Communication Quarterly*, v77 n1 p80-98
- [19] **Joseph R. Dominick,** “*Who Do You Think You are? Personal Home Pages and Self-Presentation on the World Wide Web,*” *Journalism & Mass Communication Quarterly*

- [20] **Mary Jae Paul**, (Winter 2001) "*Interactive Disaster Communication on the Internet: A Content Analysis of Sixty-Four Disaster Relief Home Pages*," *Journalism & Mass Communication Quarterly* 78: 739-753.
- [21] **Chang Liu, Kirk P. Arnett, Louis M. Capella, and Robert C. Beatty**, (1997) "*Web Sites of the Fortune 500 Companies: Facing Customers through Home Pages*," *Information & Management*: 335-345.
- [22] **Javed Mostafa**, (February 2005) "*Seeking Better Web Searches*," *Scientific American* 292:66-68.
- [23] **Xigen Li**, (Summer 1998) "*Web Page Design and Graphic use of Three U.S. Newspapers*," *Journalism & Mass Communication Quarterly*: 353-365.
- [24] **Brian L. Massey, and Li-jing Arthur Chang**, (December 2002) "*Locating Asian Values in Asian Journalism: A Content Analysis of Web Newspapers*," *Journal of Communication* 52: 987.
- [25] **William P. Cassidy**, (Winter 2005) , "*Web-only Online Sites More Likely to Post Editorial Policies Than Are Daily Paper Sites*," *Newspaper Research Journal* 26
- [26] **McTavish, G. D. & E. B. Pirro** (1990): *Contextual Content Analysis*. Quality and Quantity 24

APPENDIX

A: CODE

```
public static void FetchPage(int pageid) {
    string asama = "0";
    Console.WriteLine("Step 1");
    dbPageDefinition definition = null;
    ContentExtractor ce = new ContentExtractor();
    Console.WriteLine("Step 1.1");
    UrlAnalyzer ua = new UrlAnalyzer();
    Console.WriteLine("Step 1.2");

    int defid = 0;
    try {
        asama = "1";
        try {
            if (pageid == 0)
                defid =
Convert.ToInt32(ProjectSettings.DB0.GetValue("exec GetNewJob"));
            else defid = pageid;
            if (defid < 1) return;
            definition = new dbPageDefinition(defid);
            if (definition.Status == enPageDefinitionStatus.Working)
return;

            definition.Status = enPageDefinitionStatus.Working;
            definition.Save2Db();
            Console.WriteLine("Working on " + definition.ToString());
        } catch {
            Thread.Sleep(1000);
            return;
        }
        try {
            asama = "2";
            Console.WriteLine("Step 2");
            ua.AnalyzStartPage(definition.Link);
            asama = "3";
        } catch (Exception ex) {
            if (ex.Message.Substring(0, 5) == "UIERR") {
                int code = Convert.ToInt32(ex.Message.Substring(5));
                definition.Status = (enPageDefinitionStatus)code;
                definition.LastFetchTime = DateTime.Now;
                definition.Save2Db();
                return;
            }
        }
        asama = "4";
        Console.WriteLine("Step 3");
        foreach (string str in ua.PageLinks) {
            if (dbPage.CheckWeHave(str)) continue; //Burda
LinkUniquer ile kontrol ediliyor.
            try {
                naHtmlPage newPg = new naHtmlPage(str);
                if (newPg.Source.StartsWith("<?xml") continue;
                ce.Pages.Add(newPg);
                if (ce.Pages.Count > 50) break;
            } catch { }
        }
        asama = "5";
    }
}
```

```

        Console.WriteLine("Step 4");
        if (ce.Pages.Count > 0) {
            if (ce.Pages.Count < 5) {
                DataTable dt = new DataTable();
                ProjectSettings.DB1.FillTable(dt, "SELECT TOP " + (10
- ce.Pages.Count).ToString() + " * FROM Pages WHERE PageDefinitionID=@pdid",
                new DBParam("@pdid", definition.DBID));
                foreach (DataRow dr in dt.Rows) {
                    ce.Pages.Add(new naHtmlPage(new
dbPage(Convert.ToInt32(dr["OID"]))););
                }
            }
            if ((ce.Pages.Count < 5) && (DateTime.Today.Year>2010)) {
                DataTable dt = new DataTable();
                ProjectSettings.DB1.FillTable(dt, "SELECT TOP " + (10
- ce.Pages.Count).ToString() + " * FROM OldPages WHERE
PageDefinitionID=@pdid",
                new DBParam("@pdid", definition.DBID));
                foreach (DataRow dr in dt.Rows) {
                    ce.Pages.Add(new naHtmlPage(new
dbPage(Convert.ToInt32(dr["OID"]))););
                }
            }
            asama = "6";
            Console.WriteLine("Step 5");
            ce.AnalyzePages(definition.TitleType);
            asama = "7";
            int savedCount = 0;
            Console.WriteLine("Step 6");
            foreach (naHtmlPage pg in ce.Pages) {
                if ((!pg.AlreadyFetched) && (pg.Content.Length > 70)
&& (!pg.Content.Contains("server error"))) {
                    if (pg.Content.Contains("error: page cannot be
displayed")) continue;
                    if (pg.Content.Contains("you have an error in
your sql ")) continue;
                    if (pg.Content.Contains("error_default_404"))
continue;
                    if (pg.Content.Contains("files search results:"))
continue;
                    if (pg.Content.Contains("this domain is parked"))
continue;
                    if (pg.Content.Contains("database error:"))
continue;
                    if (pg.Content.Contains("database error:"))
continue;
                    if (pg.Content.Contains("warning: require()"))
continue;
                    if (dbPage.CheckFromDB(pg.Link)) continue;
                    if (dbPage.CheckContentFromDB(pg.Content,
pageid)) continue;
                    dbPage newPage = new dbPage(pg,
                    asama = "7.2";
                    newPage.InsertToDB();
                    asama = "7.3";
                    newPage.InsertDetails(pg);
                    Thread.Sleep(100);
                    savedCount++;
                }
            }

```

```

    }
    TimeSpan fark =
DateTime.Now.Subtract(definition.LastFetchTime);
    dbSysLog.WriteALog(definition.ToString() + " " +
savedCount.ToString() + " Yeni Haber Alındı (in " +
((int)fark.TotalMinutes).ToString() + " min.)");
    } else {
        TimeSpan fark =
DateTime.Now.Subtract(definition.LastFetchTime);
        dbSysLog.WriteALog(definition.ToString() + " Yeni Yok
(in " + ((int)fark.TotalMinutes).ToString() + " min.)");
        Console.WriteLine("There isn't any New Articles");
    }
    asama = "8";
    definition.Status =
Usishi.Projects.NetAjan.NaLib.enPageDefinitionStatus.Ready;
    definition.LastFetchTime = DateTime.Now;
    try {
        definition.Save2Db();
    } catch {
        Thread.Sleep(1000);
        definition.Save2Db();
    }

} catch (Exception ex) {

    Guru.InformException(new ObjectException("Tahmin Edilemeyen
Hata ASAMA[" + asama + "] ", ex));
    Thread.Sleep(1000);
    CrawlingError.WriteAnError(defid, ex.ToString());
    if (definition != null) {
        definition = null;
        definition = new dbPageDefinition(defid);
        definition.Status =
NaLib.enPageDefinitionStatus.ErrorReported;
        definition.LastFetchTime = DateTime.Now;
        definition.Save2Db();
    }
}

}

public void AnalyzStartPage(string url) {
    naHtmlPage mainPage = new naHtmlPage(url);
    if (mainPage.Source == null) {
        throw new Exception("UIERR"+
((int)enPageDefinitionStatus.PageTimeout).ToString("000"));
    }
    PagesDomain = GetUrlDomain(url);

    ArrayList links = new ArrayList();
    links.AddRange(LinkFinder.Find(mainPage.Source));

    PageLinks.Add(mainPage.Link);

    ArrayList heArr = new ArrayList();
    heArr.AddRange(HatedExtentions.Split('|'));

    foreach (LinkItem li in links) {
        try {

```



```

        return list;
    }
}

private void FetchPageSource(object obj) {
    string url = (string)obj;
    try {
        HttpWebRequest request =
(HttpWebRequest)WebRequest.Create(url);
        request.UserAgent = "Mozilla/4.0 (compatible; MSIE 8.0;
Windows NT 6.0)";
        request.Timeout = 30000;
        request.AllowAutoRedirect = true;
        _Response = (HttpWebResponse)request.GetResponse();
        if (!_Response.ContentType.Contains("text/html")) {
            throw new Exception("UIERR" +
((int)enPageDefinitionStatus.ContentTypeDifferent).ToString("000"));
        }

        Console.WriteLine("status code:
{0},contenttype:{1},length:{2}", _Response.StatusCode, _Response.ContentType,
_Response.ContentLength);

        PageEncoding = Encoding.GetEncoding("windows-1254");
        using (StreamReader reader = new
StreamReader(_Response.GetResponseStream(), PageEncoding)) {
            Source = reader.ReadToEnd();
        }
        _Response.Close();
    } catch {
        Source = "hatalı";
    }
}

public naHtmlPage(string url) {
    AlreadyFetched = false;
    try {
        if (url.Length < 5) throw new Exception("UIERR"+
((int)enPageDefinitionStatus.AdressIsEmpty).ToString("000"));

        Link = url;

        url = url.Substring(url.LastIndexOf("http://"));
    } catch {
        throw new Exception("UIERR" +
((int)enPageDefinitionStatus.DefinitionIncorrect).ToString("000"));
    }

    Thread th = new Thread(FetchPageSource);
    th.Start(url);
    int iTimeOut = 0;
    while (th.ThreadState == ThreadState.Running) {
        Thread.Sleep(1000);
        iTimeOut++;
        if (iTimeOut > 60) {
            _Response.Close();
            th.Abort();
            WebClient wc = new WebClient();
            Source = wc.DownloadString(Link);
        }
    }
}

```

```

        if (
            (Source.Contains("Ä±"))
            ||
            (Source.Contains("ÃŸ"))
            ||
            (Source.Contains("Ä±"))
        ) {
            if ((!Source.Contains("ö")) && (!Source.Contains("ğ")) &&
                (!Source.Contains("ı")))
                Source = Encoding.GetEncoding("utf-
8").GetString(PageEncoding.GetBytes(Source));
        }
        public override string ToString() {
            return Title + "[" + Link + "];
        }
    }

public void AnalyzePages(enTitleType pagesTitleType) {
    _RemoveCommentsConvertQuotas();
    _RemoveMetasAndCorrectHtmlTags();
    _RemoveWithRegex();
    _RemoveWithReplace();
    _RemoveTitles();
    _RemoveEmptyLines();

    string[] sameLines = _FindSameLinesInPages();
    foreach (naHtmlPage pg in Pages) {
        foreach (string str in sameLines) {
            if (str.Length>0)
                pg.Content = pg.Content.Replace(str + _splitstr, "");
        }
    }
    _RemoveCountersAndCurrencies();
    _RemoveSenselessLines();
    _RemoveExtraSpaces();

    foreach (naHtmlPage page in Pages) {
        string str = page.Content;
        if (str.Length > 10) {
            foreach (Corrections c in CorrectList) {
                str = str.Replace(c.FindStr, c.CorrectWith);
            }
            page.Content = Regex.Replace(page.Content.Replace("-",
"!TIRE!"), "[^A-Za-z0-9ŞşĞğİıÜüÖöÇç\r\n?$/?.!%\"';:€&_ ]",
"!TIRE!"), "-");
            page.Content = Regex.Replace(page.Content, "((([0-
1][0-9]))(:([0-5][0-9]))(:([0-5][0-9]))|((([0-3][0-9]))(.[0-1][0-
9]))(.201([0-9]))", "");
            page.Content = Regex.Replace(page.Content, "(\\r\\n([0-
1][0-9]))(:([0-5][0-9])\\r\\n)", "\\r\\n");
            if (
                (page.Content.Contains("t?rkiye"))
                ||
                (page.Content.Contains("a??klama"))
                ||
                (page.Content.Contains("?leti?im"))
                ||
                (page.Content.Contains("tur?zm"))
                ||
                (page.Content.Contains("tart??ma"))
            )
        }
    }
}

```



```

    ) {
        page.Content = Encoding.GetEncoding("windows-
1254").GetString(Encoding.UTF8.GetBytes(page.Content));
    }
    if (pagesTitleType == enTitleType.FindInBody) {
        string cleanContentForTitle = page.Content;

        foreach (TitleRemoves rem in TitleRemoveList) {
            cleanContentForTitle =
cleanContentForTitle.Replace(rem.RemoveString, "");
        }

        string[] words = cleanContentForTitle.Split(' ');
        page.Title = "";
        for (int i = 0; i < 7; i++) {
            if (i + 1 > words.Length) break;
            page.Title += words[i] + " ";
        }
        page.Title = page.Title.Replace("/", "
").Replace(" ", " ");
    }
    if (page.Title.Length > 80) page.Title =
page.Title.Substring(0, 80);
    } else { page.Title = " "; }
}

private string[] _FindSameLinesInPages() {

    if (Pages.Count < 2) {
        return new string[0];
    }

    //metin satırlara bölünüyor
    ArrayList worklist = new ArrayList();
    foreach (naHtmlPage page in Pages) {
        string[] strArr = page.Content.Split(_splitchr);
        worklist.AddRange(strArr);
    }

    ArrayList results = new ArrayList();
    int num2 = (int)Math.Round(((double)((double)Pages.Count) /
2.5)); // %40

    foreach (string str in worklist) {
        int tekrar = 0;
        foreach (string str2 in worklist) {
            if (str2.Trim() == str) {
                tekrar++;
            }
            if (tekar > num2) {
                if (!results.Contains(str.Trim()))
                    results.Add(str.Trim());
                break;
            }
        }
    }

    return (string[])results.ToArray(typeof(string));
}

```