A TEMPLATE-INDEPENDENT CONTENT EXTRACTION APPROACH FOR NEWS WEB PAGES

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING

AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Ahmet Yeniçağ

September, 2012

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

Prof. Dr. Fazlı Can(Advisor)

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

Prof. Dr. Özgür Ulusoy

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

Assist. Prof. Dr. Seyit Koçberber

Approved for the Graduate School of Engineering and Science:

Prof. Dr. Levent Onural
Director of the Graduate School

ABSTRACT

A TEMPLATE-INDEPENDENT CONTENT EXTRACTION APPROACH FOR NEWS WEB PAGES

Ahmet Yeniçağ
M.S. in Computer Engineering
Supervisor: Prof. Dr. Fazlı Can
September, 2012

News web pages contain additional elements such as advertisements, hyperlinks, and reader comments. These elements make the extraction of news contents a challenging task. Current news content extraction (NCE) methods are usually template-dependent. They require regular maintenance, since news providers frequently change their web page templates. Therefore, there is a need for NCE methods that extract news contents accurately without depending on web page templates. In this thesis, a template-independent News content EXTraction approach, called N-EXT, is introduced. It first parses a web page into its blocks according to the HTML tags. Then, it examines all blocks to detect the one that contains the major part of the news content. For this purpose, it assigns weights to the blocks by considering both their textual sizes and similarities to the news title. For quantifying the importance of these two weight components, we use the k-fold cross validation approach; and for assessing the impact of different possible similarity measures, we use a one-way Analysis of Variance (ANOVA) with a Scheffé comparison. The block with the highest weight is considered as the news block. Our approach eliminates the sentences in the news block that are not related to the news content by considering similarities of sentences to the news block. Finally, it also examines other blocks to detect the rest of the news content. The experimental results show the accuracy and robustness of our method by using two test collections whose web pages are obtained from several different news websites.

Keywords: Information extraction, news block detection (NBD), news content extraction (NCE), news portal, web information aggregators, wrappers.

ÖZET

HABER İNTERNET SAYFALARI İÇİN ŞABLON-BAĞIMSIZ İÇERİK ÇIKARTMA YÖNTEMİ

Ahmet Yeniçağ
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Prof. Dr. Fazlı Can
Eylül, 2012

İnternet haber sayfaları, reklamlar, bağlantılar, ve kullanıcı yorumları gibi fazladan elemanlar içermektedirler. Bu elemanlar, haber içeriklerinin çıkartılmasını zorlu kılmaktadırlar. Günümüzdeki haber içeriği çıkartma (HİÇ) yöntemleri genellikle şablon bağımlı olarak çalışmaktadırlar. Haber sağlayıcılar, internet sayfası şablonlarını sıklıkla değiştirdikleri için, bu yöntemler düzenli bakım gerektirmektedirler. Bu nedenle, haber içeriklerini internet sayfası şablonlarına bağımlı olmaksızın doğru bir şekilde çıkartabilecek HİÇ yöntemlerine gereksinim duyulmaktadır. Bu tez çalışmasında, bir şablon bağımsız haber içeriği çıkartma yöntemi (N-EXT) önerilmiştir. N-EXT ilk olarak, bir haber sayfasını HTML etiketlerine göre bloklara ayrıştırır. Daha sonra haber içeriğinin çoğunluğunu ya da tamamını içeren bloğu tespit etmek için ayrıştırdığı tüm blokları inceler. Bu amaçla, bloklara metinsel boyutlarını ve haber başlığına olan benzerliklerini göz önünde tutarak birer ağırlık tahsis eder. Bu iki ağırlık bileşenlerinin önemini belirlemek için k-kat çapraz doğrulama yaklaşımı ve olası farklı benzerlik ölçülerinin etkilerini değerlendirmek için de tek yönlü varyans analizi (ANOVA) ve Scheffé çoklu karşılaştırma testi birlikte kullanılmıştır. En yüksek ağırlığa sahip blok, haber bloğu olarak düşünülür. Haber bloğu içerisinde yer alan fakat haber içeriğiyle ilgisi olmayan cümleler, önerilen yöntem tarafından haber bloğuna olan benzerlikleri değerlendirilerek haber bloğundan elenir. Son olarak, önerilen yöntem olası haber içeriği kalıntılarını tespit etmek için, haber bloğu dışındaki blokları da inceler. Farklı haber sitelerinin internet sayfalarını içeren iki farklı deney koleksiyonu üzerinde yapılan deneylerce, önerilen yöntemin doğruluğu ve dayanıklılığı gösterilmiştir.

Anahtar sözcükler: Bilgi çıkartma, haber bloğu tespiti (HBT), haber içeriği çıkartma (HİÇ), haber portalı, internet bilgi kümeleyicileri, sarmalayıcılar.

Acknowledgement

I am deeply grateful to my supervisor Prof. Dr. Fazlı Can, who helped and guided me with his invaluable pointers in all stages of my life. I would like to thank him for giving me the opportunity to work with him for three precious years, and for his endless patience in this study.

I am grateful to the members of the jury, Prof. Dr. Özgür Ulusoy and Assist. Prof. Dr. Seyit Koçberber for using their precious times to read this thesis and giving their valuable comments about it.

I would like to address my special thanks to Dr. Kıvanç Dinçer, the vice president of Scientific and Technical Research Council of Turkey (TÜBİTAK), for his precious interests in this thesis.

I would like to acknowledge TÜBİTAK for their support under the grant number 111E030, and BİDEB for their scholarship under the program number 2210. I also would like to thank Bilkent Computer Engineering department for their financial and educational support during my studies.

I am also grateful to all of my friends, but especially to Erkam Akkurt, Sefa Alemdaroğulları, Umut Çayıröz, Emir Gülümser, Emre Gürbüz, Ceyhun Karbeyaz, Yasin Kavak, Selçuk Kızılırmak, Seçkin Okkar, Muhammed Ali Sağ, Mete Sünsüli, Kemal Şen, Mustafa Tekin, Çağrı Toraman, and Mustafa Yücefaydalı for their friendships and supports during my studies.

to my beloved family...

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

There is a dramatic increase in the amount of information on the web [1] and news constitutes a significant part of it. PRC [2] and *The Economist* [3] indicate that a large number of web users prefer reading news from news websites rather than traditional printed media. Besides, almost all news websites use news RSS (Rich Site Summary) feeds to distribute their news to the web users. RSS is an XML-based web feed format for delivering frequently changing or updated web contents such as news. It allows web users to keep track of the latest news as soon as they are published.

Current news web pages usually contain three textual news content elements: news title, news description, and news text. However, news web pages usually contain other elements such as textual and visual advertisements, links to the other websites or other web pages in the same news website, web page menus and navigation bars, comment fields, and so on. General structure of a news web page is shown in Figure 1.1: blocks labeled with letters A, B, and C are the news content elements, such that block A is the title of the news, block B is the description of the news, and block C is the text of the news; blocks D and H represent the advertisements of the web page; block E is the field where readers can write their comments about the news; and blocks F and G contain hyperlinks

to the other web pages and articles of the news website. Besides, although it is not the interest of this thesis, block I is the media file of the news. (The interest of this thesis is only textual contents of news.) These elements are not related to the news content, but together with news content elements, they constitute the template of a news web page that gives web users a more enhanced browsing experience on that news website. On the other hand, these noises make news web pages less structured and increase their heterogeneity and complicate the extraction of news content from them.

1.2 Problem Statement

Extraction of news content from news web pages is an crucial and difficult task [4]. As it is known and also confirmed by our bitter experience, it directly affects the performance of information retrieval and various web mining modules of news aggregators including indexing, ranking, web page clustering, classification, summarization, duplicate detection, new event detection, topic tracking, etc. The task that we undertake in this study follows our research group's earlier studies on information retrieval [5], new event detection and topic tracking [6], novelty detection [7], text summarization [8] and duplicate detection [9]. We employ the results of these studies and this current study in a coordinated way for the implementation of a news aggregator [10] and [11]. If news content is not extracted from news web page accurately, performance of the aforementioned modules is negatively affected. The research presented in this study is a contribution in this direction: we use news content extraction (NCE) in our news portal, called *Bilkent News Portal* [10], which uses RSS feeds to gather news web pages from various different news websites, extracts news contents from these news web pages, and displays the contents to the web users as it is seen in Figure 1.2. Bilkent News portal also uses extracted news contents in web mining modules. Thus, extracted news contents need to be noise-free so that performance of other modules used in this portal is not negatively affected. The results of our study can be used by other researchers and practitioners in their studies and information aggregations systems.

Figure 1.1: General structure of a news web page.

Figure 1.2: Main page of Bilkent News Portal.

1.3 Wrappers and Their Problems

Most of the traditional methods manually or automatically generate wrappers to extract the news content from web pages. Wrappers perform web page content extraction by recognizing the template of web pages. Liu [12] indicates that since they are template-dependent, due to this property in general they only work for the web pages that they are generated for. These approaches need to be trained on a set of manually labeled samples before they can be used in the extraction process. However, web pages of different news websites have different templates, which require a modification in the approach or training for each different web page template. But training the approaches for each different web page template or modifying the approach with respect to any change in the template is costly, inefficient, and most importantly not automatic. Therefore, an extraction method needs to be robust and generic, such that it has to extract the news content accurately without depending on the web page templates.

Han et. al [13] state that traditional wrapper-based web page content extraction approaches need considerable maintenance to work properly for a long period of time, which is a difficult and costly work, since templates change frequently. Vadrevu et. al [14] specify that wrapper-based approaches also need human intervention, since manually labeled web pages are required by these approaches to learn the template of websites. However, Arasu et. al [15] indicate that human input is time consuming and error-prone. Additionally, some methods try to automatically detect the template of the news web pages; however, these methods are less accurate if the number of web pages analyzed to detect the template is not large enough [13]. Web page templates change frequently; therefore, providing large number of pages to feed the template detection method is mostly problematic.

1.4 Proposed NCE Approach: N-EXT

In this thesis, we propose an automatic template-independent web News content EXTraction approach, called N-EXT, which uses blocking tags to parse a news

web page into blocks, and extracts the news contents from these blocks. The major part of the news text is stored in one of the blocks, and it is referred to as the news block. Detecting the news block in a template-independent content extraction approach is a critical step in the extraction process. If the news block is not detected correctly news content extraction accuracy decreases. Ziyi et. al [16] uses largest block approach, which considers only number of words in blocks to detect the news block. But our experiments show that this approach is not accurate enough. For this purpose, we propose a news block detection (NBD) approach, which assigns weights to blocks by considering both their textual size and similarity to news title. The one with the highest weight is considered as the news block. We use an HTML parser to generate Document Object Model (DOM) tree of the web page, and treat all nodes represented with current blocking tags as blocks rather than trying to detect the blocking tag of a web page as it is done in the largest block approach. (The largest block approach determines the frequencies of candidate blocking tags, <DIV> and <TABLE>, in a web page and selects the one with the highest frequency as the blocking tag, and divides the page into blocks according to the selected tag.) The experimental results show that our proposed NBD approach outperforms the largest block approach and can be used in practical environments due to its high NCE accuracy.

As will be illustrated in detail later, N-EXT first parses an HTML news web page to identify its blocks according to the HTML tags. Then, it detects the news block that contains the news content by ranking the web page blocks according to both their textual size and similarity to the news title. It eliminates the sentences in the news block that are not related to the news content by calculating similarities of sentences to the news block. It examines other blocks to detect the rest of the news content if any exists.

1.5 Research Contributions

In this study, we

- Propose an NCE method (N-EXT) that extracts news contents accurately

without depending on the web page templates, and does not require any regular maintenance or human intervention,

- Demonstrate the robustness of our method by showing its sustained success in different environments,
- Outperform the largest block approach by considering not only block size but also block similarity to the news title,
- Show the positive impact of removing the hyperlink texts from blocks on the detection of the news block,
- Show that stemming improves the content extraction accuracy,
- Provide an NCE test collection, which also incorporates an NBD component, for news content extraction that we will share with other researchers; to the best of our knowledge there is no previous standard NCE test collection.

1.6 Overview of the Thesis

The rest of the thesis is organized as follows. Chapter 2 gives an overview about existing content extraction approaches by categorizing them according to techniques they use for content extraction. Chapter 3 provides background information about this study. Chapter 4 introduces our proposed web NCE method (N-EXT) in terms of the stages involved. Chapter 5 defines the measures that will be used to evaluate the performance of the proposed NCE approach. Our Turkish and English test collections are described in Chapter 6. Besides, the experimental results with their evaluations are also given in Chapter 6. Chapter 7 gives configuration information about Bilkent News Portal. Finally, we conclude with a summary of our findings, and provide some future research pointers.

Chapter 2

Related Work

Chang et. al [4] consider the problem in a more general web page information extraction (IE) point of view, provide a comprehensive survey, and indicate that the extraction target of an IE task can be a relation of k-tuple (k is the number of attributes in a record) or it can be a complex object with hierarchically organized data. They compare IE systems in three dimensions: a) the "task domain" that aims to explain why a system fails to handle some websites of particular structures, b) the "automation degree" that aims to classify systems based on the techniques used, and c) the "technique used" that aims to measure the degree of automation for such systems.

Until today, numerous researches have been done, and researchers tried to find methods for extracting the information from web pages automatically, and accurately. Earlier works were generally semi-automatic information extraction approaches, which generate wrappers to extract information. Then, automatic information extraction approaches have taken the place of these semi-automatic approaches. In the following sections, an overview of existing semi-automatic and automatic information extraction approaches is introduced. Summaries of the approaches presented in the following sections are presented in Tables 2.1 and 2.2

2.1 Wrapper-based Approaches

Most of the traditional information extraction approaches manually or automatically generate wrappers to extract news contents from web pages [12]. Wrappers perform content extraction from web pages by recognizing templates of web pages. Some of the existing information extraction approaches that generate wrappers to extract contents from web pages are classified as semi-automatic, since they need to be trained on a set of manually labeled samples before they can be used in the extraction process. Although many of the wrapper-based approaches are semi-automatic, there are also some automatic approaches.

Laender et. al [17] present a taxonomy, which is based on the methods used by information extraction approaches to generate wrappers, and provide a quantitative analysis of them. They categorize existing manual, semi-automatic, and automatic approaches into six groups with respect to the method they used for wrapper generation: 1) declarative languages-based, 2) HTML structure analysis-based, 3) Natural Language Processing (NLP)-based, 4) machine learning-based, 5) data modeling-based, and 6) ontology-based. In the following subsections, five of these six groups are explained with details of their representative approaches.

2.1.1 Declarative Language-based Wrappers

Some programming languages, which are alternative to commonly used ones in wrapper generation such as Java, are developed in purpose to help researchers in generating wrappers. These languages are specific to the wrapper generation task. One of the best known approaches, which use languages declared for wrapper generation, is *WebOQL* [18]. Other approaches that develop languages for wrapper generation are *Minerva* [19], *TSIMMIS* [20], *Jedi* [21], and *FLORID* [22].

Arocena and Mendelzon [18] propose a query-like language, called *WebOQL*, which is declared for extracting data from HTML web pages. WebOQL has two main components: the data model and the query language. WebOQLs data model considers the web as a graph of tree. It parses an HTML web page into a special

kind of ordered tree, called *hypertree*. Users can search a piece of information in the hypertree by writing queries. WebOQL's query language returns the result of the query by navigating through the hypertree to locate the information queried.

2.1.2 HTML Structure Analysis-based Wrappers

HTML web pages have structural features such that they are organized by HTML tags. Some of the information extraction approaches uses these structural features of HTML web pages for generating wrappers to extract information. These approaches parse HTML web pages into trees with respect to their HTML tags, and generate extraction rules to detect templates of the web pages, such as RoadRunner [23]. Some other approaches based on the structural features of HTML web pages are *W4F* [24], and *XWRAP* [25].

Crescenzi et. al [23] propose an IE approach, called *RoadRunner*, which uses the structural features of HTML web pages to automatically generate wrappers for information extraction. A sample set of web pages from the same website are compared to generate an extraction rule based on the differences and similarities between them. Each extraction rule is generated for a specific website and can deal with only HTML web pages of that website. Relevant information is extracted from the HTML web pages using the generated extraction rules.

2.1.3 Natural Language Processing (NLP)-based Wrappers

Some information extraction approaches use natural language processing (NLP) techniques such as part-of-speech (POS) tagging to generate wrappers. These approaches use NLP techniques to learn pattern-match extraction rules by generating semantic constraints that are used to detect the relevant information within a document containing only textual information. RAPIER [26] is one of the most popular IE approaches that use NLP techniques for wrapper generation. There are also some other approaches that use NLP-based wrappers such as *SRV*, [27], and *WHISK* [28].

Califf and Mooney [26] propose an IE approach, called *RAPIER (Robust Automated Production of Information Extraction Rules)*, which uses NLP techniques to extract information from natural language documents that contain only textual information written in natural languages. RAPIER requires a filled template, which represents structure of the information to be extracted. It uses that template to learn extraction pattern-match rules. Each extraction rule consists of three parts: 1) a pre-filler pattern that specifies the text exactly before the filler, 2) a pattern that specifies the actual slot filler, and 3) a post-filler pattern that specifies the text exactly after the filler. Each pattern matches only a single word or symbol from each document. Pattern-match rules extract the fillers from the documents for the slots in the template.

2.1.4 Machine Learning-based Wrappers

Information extraction approaches, which use machine learning techniques for wrapper induction, generate extraction rules to extract information similarly with the approaches that use NLP techniques. Although, both techniques generate delimiter-based extraction rules, which means they specify patterns exactly before and after the text to be extracted in the document; however, approaches which use machine learning techniques that rely on the features that specify the structure of information to be extracted rather than the linguistic constraints NLP-based approaches rely on. *STALKER* [29], *W1EN* [30], *SoftMealy* [31], and the approach proposed by Zheng [32] are representatives of the approaches that use machine learning techniques.

Muslea et. al [29] propose a wrapper induction approach, called *STALKER*, which uses machine learning techniques to generate rules for IE. Before the rule generation process, user needs to provide a labeled set of training samples by using the graphical user interface (GUI) offered by the approach to mark up the relevant information in the samples. GUI generates sequences of tokens which represent the start rules (prefixes) of the information to be extracted from the marked samples. STALKER generates an extraction rule from these generated sequences of tokens. If sequences of tokens do not match with each other, which means

samples do not share a common template, STALKER generate an extraction rule for each pattern, and returns a set of extraction rules. These rules are used to extract relevant information from the documents.

2.1.5 Data Modeling-based Wrappers

Some of the information extraction approaches generate a data model that represents the structure of the web pages or the plain text files from where relevant data is extracted. Data modeling primitives, such as trees or lists, which consist of nodes or elements that represent the structural components of the documents, are used for generating the data model. After modeling the data source, these approaches try to locate the relevant information in the model by generating extraction patterns similarly with NLP-based and machine learning-based approaches. Approaches that adopt data modeling are *NoDoSE* [33] and *DEByE* [34].

Adelberg [33] propose an IE approach, called *NoDoSE (Northwestern Document Structure Extractor)*, to extract information from documents by determining their structures. NoDoSE requires labeled samples from users. Thus, it offers to users a GUI, which is used to decompose the document to identify the data of interest. Then, NoDoSE maps the decomposed document into a document tree. Each node of the tree represents one of the structural components of the document such as a record of a list, which holds the starting and ending offset values indicating the portion of the document that corresponds to the relevant data. NoDoSE infers the structure of the document from the tree, and extracts the relevant data.

IE Method	Work	Degree of Automation	Advantages	Disadvantages
Declarative language-based wrapper	Arocena and Mendelzon approach (WebOQL) [18]	Manual	(a) allows the representation of objects with structural variations	(a) user must examine the web pages and find the HTML tags that separate the objects of interest (b) require the user to execute all the wrapper generation process manually (c) works only for HTML data sources
HTML structure analysis-based wrapper	Crescenzi et. al approach (RoadRunner) [23]	Automatic	(a) allows the representation of objects with structural variations (b) does not require any user intervention, besides providing sample pages (c) easy to use	(a) works only for HTML data sources (b) extraction rules generated are specific to websites
NLP-based wrapper	Califf and Mooney approach (RAPIER) [26]	Semi-automatic	(a) good for information extraction from natural language documents	(a) user must provide training samples (b) does not support objects with structural variations
Machine learning-based wrapper	Muslea et. al approach (STALKER) [29]	Semi-automatic	(a) requires fewer samples (b) allows the representation of objects with structural variations (c) offers a GUI to users for marking up the relevant information in the samples	(a) user must provide labeled samples (b) extraction rules generated are specific to websites
Data modeling-based wrapper	Adelberg approach (NoDoSE) [33]	Semi-automatic	(a) offers a GUI to users for decomposing the samples (b) allows the representation of objects with structural variations (c) supports a variety of formats to output the data extracted	(a) user must provide labeled samples

Table 2.1: Overview of existing wrapper-based information extraction approaches.

2.2 Classifier-based Approaches

Supervised learning is another technique that is used for information extraction. Some IE approaches treat the extraction problem as a classification task. Approaches that use supervised learning techniques generally depend on a classifier such as *Support Vector Machine (SVM)* or *Condition Random Fields (CRF).* These classifiers are trained on a set of samples before being used in the extraction process. Each part of a web page is classified as title, text, author, etc. by classifiers by using structural or semantic features , and the parts that contain relevant information are extracted from the web pages.

Ibrahim et. al [35] propose a supervised machine learning classification approach, which uses an SVM classifier to extract textual elements, titles and full text, from news web pages. Proposed approach parses an HTML web page into parts with respect to HTML tags (<DIV>, <TD>, <P>, and
). Some features, such as length of text, percentage of hypertext (the text bounded by <a> tag), percentage of meta-script text (the text bounded by <meta> and <script> tags), percentage of decoration text (the text bounded by <input>, <select>, and <option> tags), and percentage of image, are extracted from blocks, and each block is classified by using those features as a title, a full-text, or other. Parts that contain relevant information, which means they are classified as a title or a full-text, are extracted from the news web pages by the classifier after training the classifier on a set of samples.

Besides, instead of SVM classifiers, some other proposed IE approaches, such as [36], [37], and [38], use Conditional Random Fields (CRF) as classifiers for the extraction process. In addition, Spengler et. al [39] compare support vector machines (SVM) with conditional random fields (CRF) on a real-world web news content extraction task.

2.3 Heuristics-based Approaches

Rather than generating pattern-match extraction rules, some researchers define various heuristics that are used to recognize the desired information in documents. Information extraction approaches, which use heuristics for extraction, analyzes the web page or the document, and then extract the information from these sources by filtering them with respect to the heuristics that they use. Different sets of heuristics are used for recognizing different kinds of information to be extracted, such as text or image. Approaches proposed by Parapar and Barreiro [40], and Gupta and Hilal [41] adopt defining and using content extraction heuristics. Besides, Gottron [42] propose a system, called *CombinE*, to test and evaluate combinations of various existing and newly described content extraction heuristics.

Parapar and Barreiro [40] propose an IE system called, *NewsIR*, which recognizes and extracts news content elements (news title, news body, and news image) from news web pages by using the heuristics described by themselves. Different sets of heuristics are proposed to identify different parts of a news document. To detect if a web page is a news web page, and if it is a news web page, to identify and extract the news body from that news web page, they propose a set of heuristics, including that news are composed of paragraphs that are next to each other, paragraphs are mostly text, and only styling markup and hyperlinks are allowed in paragraphs, a low number of hyperlinks are allowed in paragraphs, and so on. Furthermore, they also propose a set of heuristics, which utilize domain specific characteristics, to detect news titles and news images, if they exist. According to their heuristics, news title is mostly placed on the top of news body, and has a special font style; and news image is placed after or inside the news body.

2.4 Relevance Analysis-based Approaches

Relevance between elements of web pages or documents, such as paragraphs, sentences, etc., is used to detect the desired information in these data sources.

In contradistinction to other traditional information extraction approaches, relevance analysis-based approaches do not analyze web page layouts, which is a time-costly work, before the extraction process. These approaches analyze the full text of a web page only during the extraction to extract all relevant information from that web page. Approaches proposed by Han et. al [13] and Wu et. al [43] are the representatives of those that use relevance analysis for IE.

Han et. al [13] proposed an IE approach based on relevance analysis. Proposed approach first obtains the news title from an RSS feed. Then, it gets the keyword list from the obtained news title. It uses the keywords in the list to detect the position of the news title in the news web page. Then, it makes a full analysis of the web page to detect all paragraphs of news content by using the detected news title position and the keyword list, and extracts them from the news web page.

2.5 Tree Edit Distance (TED)-based Approaches

HTML web pages have structures which can be easily represented by special trees, such as Document Object Model (DOM) tree. Some of the information extraction approaches utilize the structural feature of HTML web pages by evaluating the structural similarities between web pages of the same website. *Tree Edit Distance (TED)*, which is first introduced by Levenshtein [44], is the minimum cost of transforming one tree into another by a sequence of operations consisting of inserting new nodes, deleting and relabeling existing nodes. TED is used to calculate structural similarities between web pages. A generic representation is constituted for web pages that are structurally similar. Extraction patterns, which detect and extract the desired information, are generated from the generic representation of the web pages. Approaches proposed by Reis et. al [45] and Lan [46] use TED-based information extraction.

Reis et. al [45] propose a domain oriented IE approach, which use structural analysis of news web pages. Proposed approach map an HTML news web page

into a special type of tree, called *labeled ordered rooted tree*. TED is used to calculate structural similarities between labeled ordered rooted trees that represent news web pages of the same website. During the calculations, a cost is assigned to each of three operations: node removal, node insertion, and node replacement in the tree. With respect to the TEDs calculates, similar web pages are gathered into clusters that share common characteristics. Relying on the assumption that news content elements have common formats and layouts, a generic representation is constituted for each cluster to represent the structure of the web pages in that cluster. Then, a special kind of extraction pattern, called *node extraction pattern (ne-pattern)* is generated from the representation. The relevant information is extracted from the trees using ne-patterns.

2.6 Visual Features-based Approaches

People gain some experiences during browsing web pages, and subconsciously use these experiences while they are browsing other similar web pages. For instance, when people are browsing news web pages, they seek the part of the web page that contains news content by looking for some visual features of that part such as its area is larger than other parts around it, there is bold-faced sentence or phrase at the top of it, it consists of contiguous textual paragraphs, and so on. These visual features help users to distinguish the part containing the news content from other parts. Based on this idea, some information extraction approaches simulates how a reader grasps a web layout structure based on his visual perception, and try to utilize the visual features of web pages (layout, area size, font size and type, etc.) to extract the desired information. Approaches proposed by Zheng et. al [47] and Cai et. al [48] are representatives of those based on visual consistency of web pages.

Zheng et. al [47] propose a news content extraction approach to easily detect news contents by using visual consistency of news web pages. Proposed approach first maps a web page into a visual block tree, in which each node represents a rectangular area of that web page. During the mapping, instead of using HTML tags, a set of visual features (*position features*: left, top; *size features*: width,

height; *rich format features*: font size, font type; and *statistical features*: image count, hyperlink count, paragraph count, etc.) are used to represent each part in the web page. Then, proposed approach derives a composite visual feature, which is stable enough to represent the domain-level visual consistency. Then, it uses a machine learning technique (*Adaboost* [49]) to generate a vision-based wrapper, called *V-Wrapper*, for extracting the desired information. V-Wrapper is generated after training it on a set of manually labeled web pages.

2.7 Block-based Approaches

Some approaches use block-oriented structure of web pages for information extraction. These approaches parse web pages into functional areas, called *blocks*, with respect to some criteria, such as HTML tags. News web pages store informative contents into one or more of the blocks. However, web pages also contain several non-informative contents, such as textual and visual advertisements, links to other web pages, navigation bars, comment fields, etc. Hence, these approaches try to detect the block that contains informative content by using different techniques.

Debnath et. al [50] propose an approach to detect the content blocks in a web page by looking for 1) blocks that do not occur a large number of times across web pages, and 2) blocks with desired features (text, tag, list, and style sheet). Similarly, also Ho and Lin [51] try to discover the informative content blocks in a web page. But they detect them in another way as their proposed approach calculates the entropy value based on the occurrence of each term in a block, and dynamically selects the entropy threshold value, which determines either a block is informative or redundant. Ziegler and Skubacz [52] propose an approach , which extracts the blocks that contain news content from HTML web pages by computing linguistic and structural features for each block, and deciding whether a block is a signal or noise. Shen and Zhang [53] propose a block-level links based content extraction approach, which considers the web pages as continuous block-level text, and detects the block that contains news content by ranking blocks according to both their textual sizes and link counts.

Ziyi et. al [16] propose a news content extraction approach based on blocking tags. Proposed approach first detects the blocking tag of a web page by considering the occurrences of certain HTML tags, (<DIV> and <TABLE>). HTML tag that occurred most times in the web pages is determined as the blocking tag of that web page. Then, it divides that web page into the blocks and selects the block with the highest textual size, which means the block that contains the most number of words (terms), as the block containing news content. Finally, it extracts the news content from the selected block. This study is the most similar study to the study given in this thesis.

2.8 General Overview of Related Work

As mentioned earlier, existing content extraction approaches generally have some disadvantages. The wrappers-based approaches mostly depend on the template of the web pages, and for each different website, a wrapper is generated, which is a costly work. Besides, most of these wrapper-based approaches require a training stage or human intervention to manually label web pages. During the training stage, if the training dataset is not large enough, as expected a less accurate performance is obtained. On the other hand, extraction rules generated by the approaches mentioned above are usually specific to a website, and they need to be modified for different websites. Similarly, information extraction approaches other than wrapper-based ones also have some disadvantages: some of them require manually labeled samples; some of them get less accurate results if the provided samples are not comprehensive enough; some of them need regular maintenance; and some of them require several threshold values for the selection of visual features. Besides, detecting the block that contains the news content cannot be achieved accurately enough with block-based approaches. However, the approach proposed in this thesis is template-independent, and can be directly used for extracting contents of different websites without requiring any maintenance or human intervention. Additionally, it can detect the block that contains news content very accurately.

IE Method	Work	Degree of Automation	Advantages	Disadvantages
Classifier-based extraction	Ibrahim et. al approach [35]	Automatic	(a) high extraction accuracy with adequate number of samples (b) appropriate for news web pages that do not follow proper DOM tree standards	(a) less accurate if samples are not comprehensive enough (b) no support for non-HTML sources
Heuristics-based extraction	Parapar and Barreiro approach (NewsIR) [40]	Atuomatic	(a) high precision and recall values (b) detect news content elements other than news body (news title and news image)	(a) need regular maintenance for updating heuristics (b) no support for non-HTML sources
Relevance analysis-based extraction	Han et. al approach [13]	Automatic	(a) high precision and recall values (b) no need for a full analysis of web page layout before extraction	(a) no support for non-HTML sources (b) news title itself is not always dependable to detect news content paragraphs
TED-based extraction	Reis et. al approach [45]	Automatic	(a) simple implementation (b) describes a new highly efficient tree structure	(a) works only for structural data sources (b) accuracy results are relatively low
Visual features-based extraction	Zheng et. al approach [47]	Automatic	(a) easier wrapper maintenance (b) good extraction performance even with structural diversity	(a) requires too many thresholds that needs to be trained (b) user must provide labeled samples
Block-based extraction	Ziyi et. al appraoch [16]	Automatic	(a) has a web news search engine (b) high extraction accuracy if the block that contains news content is correctly detected	(a) considering only size during news block detection is not accurate enough

Table 2.2: Overview of other existing information extraction approaches.

Chapter 3

Background Information

3.1 Terminology

In the following, we define the basic components of news web pages.

Block: It is a small part of an HTML web page which is enclosed by blocking tags. Each block may consist of other blocks or segments.

Block Node: It is the node in a DOM tree [54], which represents a block of an HTML web page. Each block node may have block node children in a DOM tree.

Blocking Tag: It is the HTML tag, <DIV> or <TABLE>, which is used to separate the elements of a web page (such as advertisements, hyperlinks, and textual contents) from each other.

Leaf Block Node: It is the block node which has no block node children in a DOM tree. Leaf block nodes may have children nodes other than block nodes.

News Block: It is detected among all blocks within a news web page, which generally contains major part of the news content, at least the news text. News content elements other than the news text (news title and news description) may also be placed in the news block, but depending on the template of a news web page, these elements may also be placed in other blocks.

News Node: It is a leaf block node which is selected as the node that represents the news block among all leaf block nodes.

Segment: It is a small part of an HTML web page other than blocks, and enclosed with HTML tags other than blocking tags such as <P>,
, etc.

3.2 HTML News Web Pages and the DOM Tree

An HTML web page is organized by HTML tags including <DIV>, <TABLE>, <P>, etc. HTML tags divide an HTML web page into smaller parts, called blocks and segments. An example HTML news web page is shown in Figure 3.1.

As it is seen in Figure 3.1, there are totally seven blocks that are numbered from 1 to 7, and three segments in the example news web page. The block number 5 in Figure 3.1 is the news block of that web page, since it contains news content elements: the news description, and the news text. Although all news content elements are placed in a single block in the example web page given in Figure 3.1, they may be placed in more than one block in other news web pages.

DOM represents an HTML web page as a tree structure. DOM uses HTML tags of an HTML web page to define the tree structure of that web page. The DOM tree generated from the example HTML news web page is shown in Figure 3.2. Each node in the DOM tree represents a block or a segment of that news web page. News related elements are placed in one or more of these nodes. Nodes that are numbered from 1 to 7 are block nodes, and among these nodes, 3, 4, 5, 6 and 7 are the leaf block nodes.

3.3 News RSS Feeds

RSS (Rich Site Summary) is an XML-based web feed format for delivering frequently changing or updated web contents such as news. It allows web users to keep track of the latest news as soon as they are published by news websites.

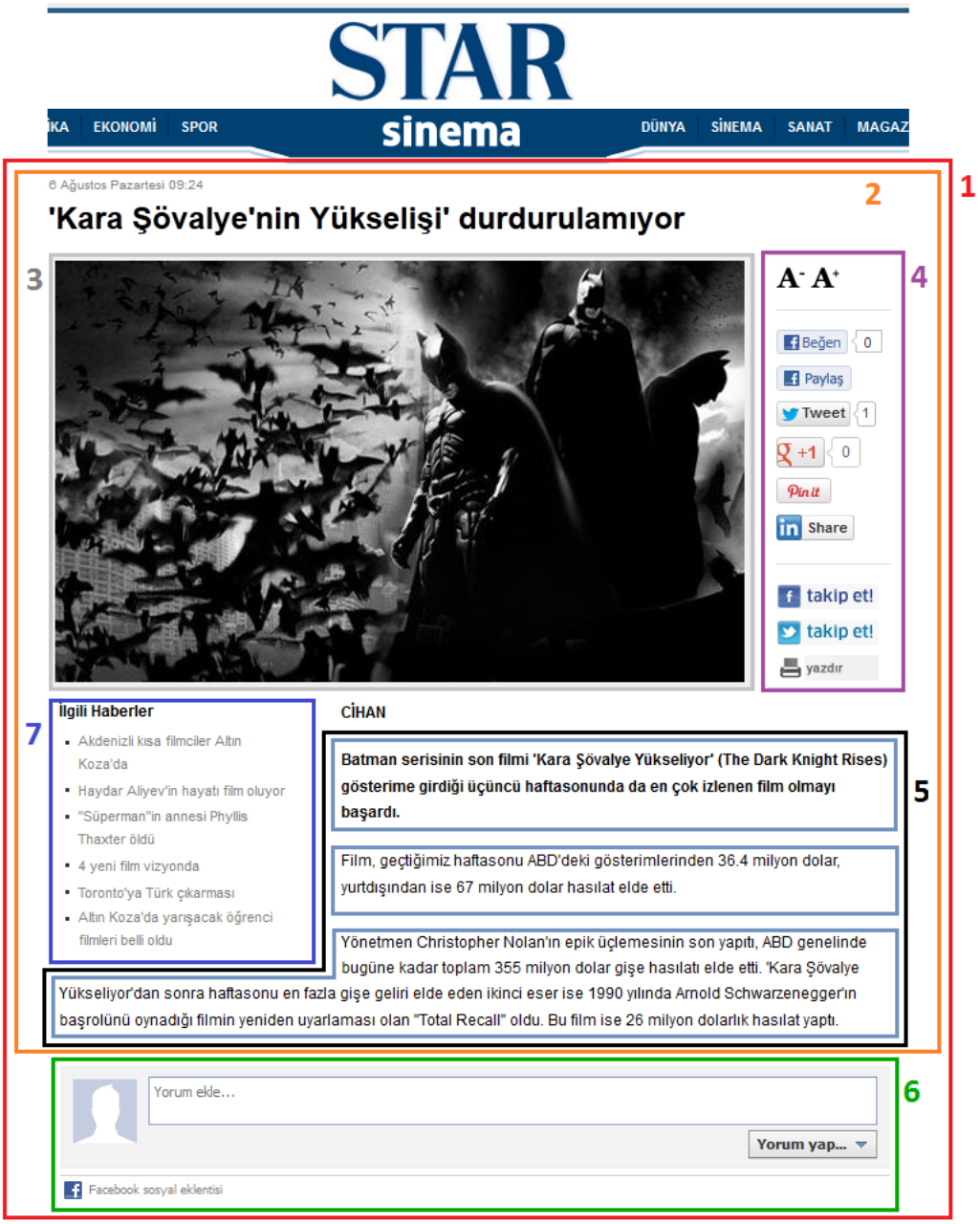


Figure 3.1: An example HTML news web page divided into its blocks/segments.

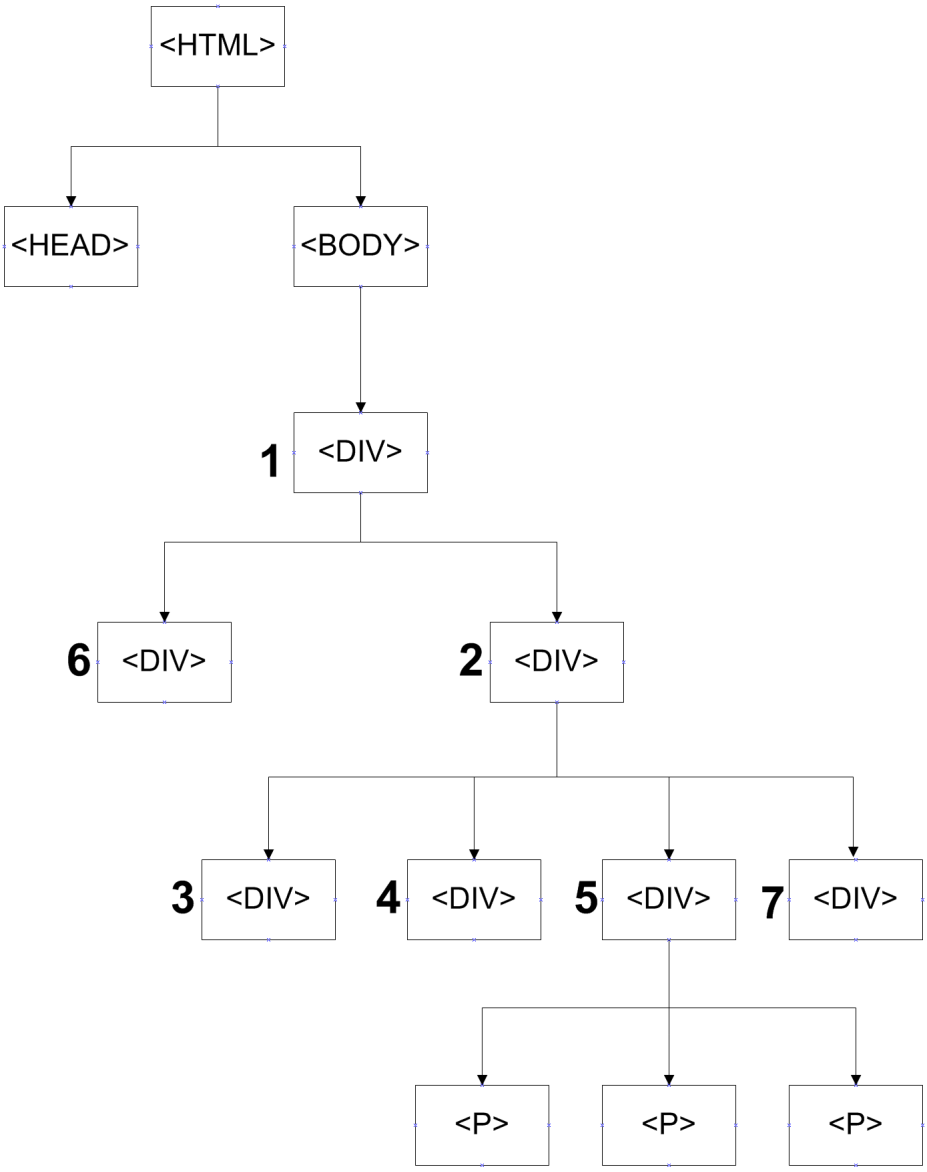


Figure 3.2: DOM tree generated from the example HTML news web page of Figure 3.1.

News RSS feed is a document, which consists of items each of which generally contains news title, brief description of the news, Uniform Resource Locator (URL) link of the news, category of the news, and publication date of the news. An example RSS feed is shown in Figure 3.3.

```
  <category domain="http://www.nytimes.com/namespaces/">Histori
  <category domain="http://www.nytimes.com/namespaces/">Murders
</item>
<item>
  <title>A Distinctly British Show Closes Olympics</title>
  <link>http://feeds.nytimes.com/click.phdo?i=6a2b0a99fea07fcef
  <pheedo:origLink>http://www.nytimes.com/2012/08/13/sports/oly
  <guid isPermaLink="false">http://www.nytimes.com/2012/08/13/s
  <atom:link rel="standout" href="http://www.nytimes.com/2012/0
  <media:content url="http://graphics8.nytimes.com/images/2012/
  <media:credit>Chang W. Lee/The New York Times</media:credit>
  <description>The host of the Summer Games capped a fortnight
  <dc:creator>By DAVID SEGAL</dc:creator>
  <pubDate>Mon, 13 Aug 2012 18:17:27 GMT</pubDate>
  <category domain="http://www.nytimes.com/namespaces/keywords/
  <category domain="http://www.nytimes.com/namespaces/keywords/
  <category domain="http://www.nytimes.com/namespaces/keywords/
</item>
<item>
  <title>IHT Rendezvous: Olympic Politics: What Does It Mean fo
  <link>http://feeds.nytimes.com/click.phdo?i=6d685735b37425c39
```

Figure 3.3: Example news RSS feed.

News websites that distribute their news via RSS feeds use different RSS feed for each different news category such as business, politics, world, health, sport, science, technology, magazine, and so on. Bilkent news portal [10] gathers news of several different news categories from several Turkish news websites by using news RSS feeds for each news category of these news websites.

Chapter 4

The Method: N-EXT

N-EXT consists of six stages: 1) parsing news RSS feed to obtain title, publication date, and URL link of the news, 2) parsing HTML news web page into blocks, 3) eliminating noises from blocks, 4) detecting the news block among all cleaned blocks, 5) extracting the news content from cleaned news block, and 6) examining other blocks to detect the rest of the news content if any exists. These stages are further explained in detail in this section. General schema of N-EXT is shown in Figure 4.1.

4.1 Stages of N-EXT

4.1.1 Parsing News RSS Feed

In this preprocessing stage, RSS feeds are parsed in order to get title, publication date, and URL link of each news document. After obtaining the URL link of a news document, HTML web page of the news document is downloaded from that URL link to be used in the NCE process. Since news RSS feeds are updated periodically, we prefer to collect news documents from news websites periodically. At the beginning of every two hours, N-EXT first updates the RSS feeds of each news website by re-downloading RSS feeds from their news websites, and repeats the procedure: parses the updated RSS feeds, obtains the URL links of latest

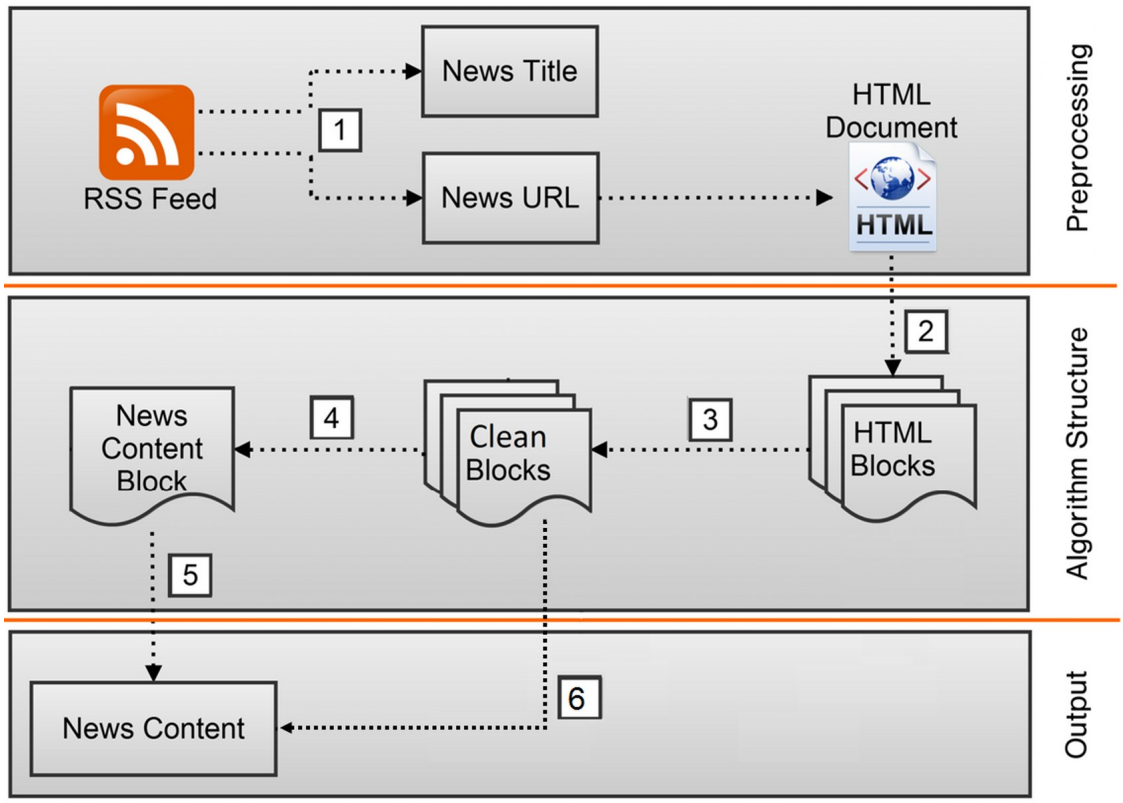


Figure 4.1: General schema of the proposed web NCE method (N-EXT).

news documents which are published in the previous two hours, and downloads HTML web pages of news documents from the obtained URL links. A list of Turkish news RSS feeds that are used by Bilkent News Portal is given in Table A.3

4.1.2 Parsing HTML Web Page

After downloading an HTML news web page, the web page is parsed into blocks and segments as it is shown in Figure 3.1, and a DOM tree is generated from it, shown in Figure 3.2, by using the Jericho HTML parser [55].

Jericho accepts an HTML web page as the input, parses the page using its HTML tags, and generates its DOM tree as the output. After parsing an HTML web page and generating its DOM tree, each node in the DOM tree has four kinds of information: 1) the HTML tag identity that encloses the block or the segment

it represents, 2) the text placed between HTML tags of the node, 3) its parent node information, and 4) the list of its children nodes. By using the methods of the Jericho HTML parser, N-EXT traverses the DOM tree generated from the HTML web page in *depth-first order* [56] to detect leaf block nodes in the DOM tree. In depth-first order, program starts from the root node and explores all successor nodes in a branch before exploring other branches.

We observed that in most cases, each piece of information is placed separately in the leaf block nodes, which represents the blocks that do not consist of any nested blocks. Although leaf block nodes are the lowest level block nodes in a DOM tree, they may consist of other segments. Segments do not contain any of news content elements as a whole. They may contain only a small part of them such as a paragraph of the news text. N-EXT aims to obtain the leaf block nodes, which contain news content elements. So, N-EXT traverses the DOM tree generated from the HTML web page and seeks the leaf block nodes in the DOM tree.

N-EXT decides whether a node in a DOM tree is a leaf block node by looking both HTML tag and children nodes of that node. Before searching the children nodes of a node, N-EXT first examines the HTML tag of that node. If the HTML tag of a node is one of the blocking tags, <DIV> or <TABLE>, N-EXT realizes that it is a block node, and it starts to traverse all its successor nodes, i.e., all nodes that are under the node itself in the DOM tree, to detect any nested block nodes. If a block node does not have any successor block nodes, then it is labeled as a leaf block node. After labeling a node as a leaf block node, N-EXT extracts the text of that node, and keeps that information in a list.

At the end of this stage, N-EXT keeps a list of leaf block nodes in the DOM tree along with the text placed between HTML tags of the nodes. Figure 4.2 demonstrates the detection process of leaf block nodes in a DOM tree.

4.1.3 Cleaning Blocks: Eliminating Noises from Blocks

After parsing the HTML web page and obtaining the leaf block nodes along with the text placed between HTML tags of the nodes, all noises which could not be

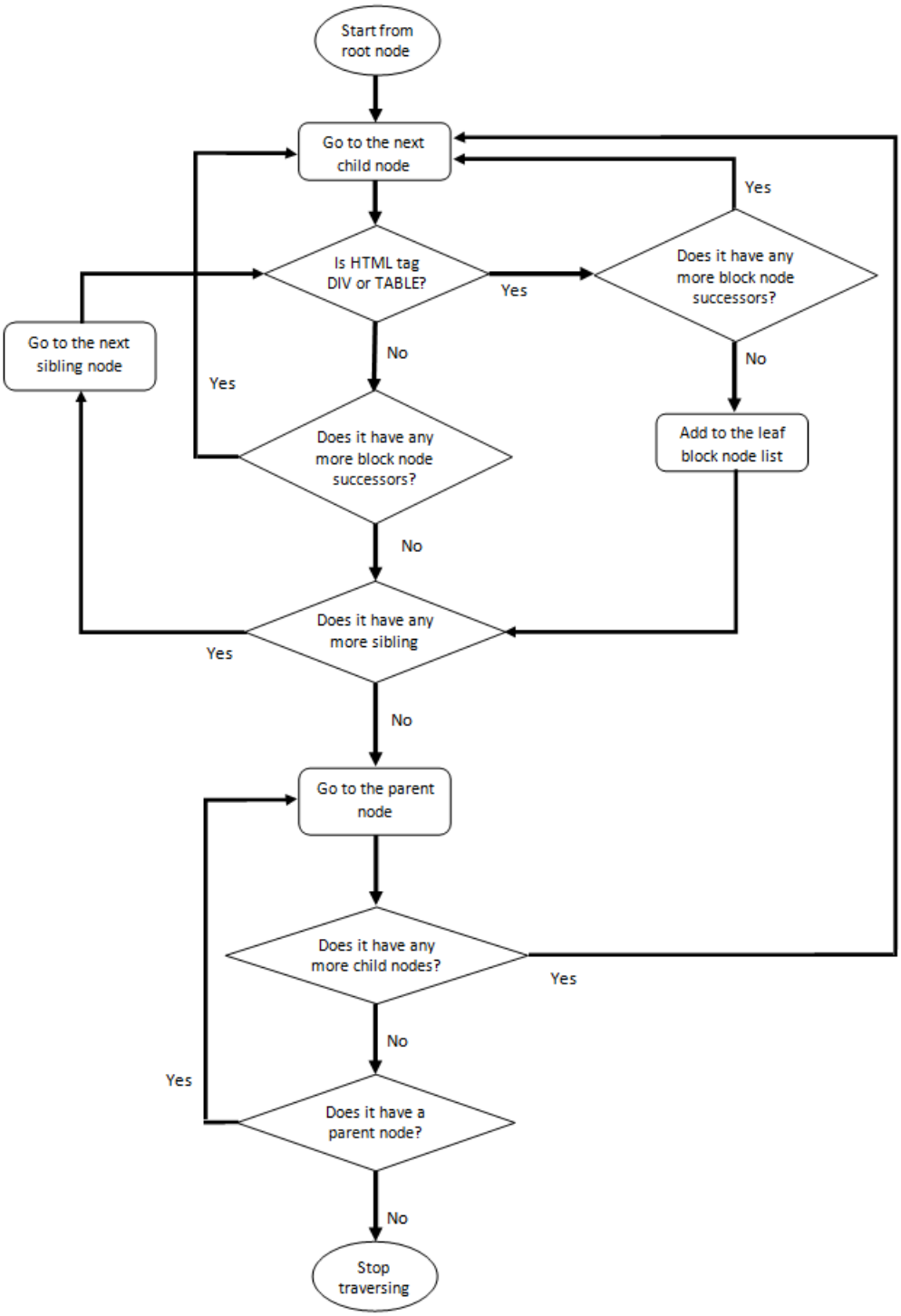


Figure 4.2: Demonstration of detecting leaf block nodes in a DOM tree.

eliminated in the previous stage due to a list of reasons are eliminated from the leaf block nodes in the cleaning stage, so that leaf block nodes contain only the text information which may or may not be news related. The reasons are:

- The HTML parser that is used to generate DOM tree, the Jericho HTML parser treats all HTML tags as a pair. For instance, if an HTML block/segment is started with "<a" tag, Jericho works with the rule that it must end with "/a>". But in HTML, there are some HTML tags which do not obey this rule such as *"input"*, *"img"*, *"iframe"*, and *"link"*. These HTML tags end with just "/>", so Jericho accepts these tags as regular texts, not HTML codes, and could not eliminate them in the previous stage. Therefore, N-EXT eliminates these tags in the cleaning stage.
- Almost all news web pages contain hyperlinks, which are references to another web pages. The size of the texts containing hyperlinks sometimes becomes a problem, since N-EXT looks at the textual size of the blocks to detect the news block. But, hyperlinks are not actually related to the news content of the current news web page, they are only references to other web pages. Therefore, N-EXT eliminates hyperlink texts, which are enclosed by "<a>" tag, to get better news block detection (NBD) accuracy.

4.1.4 Detecting the News Block Using Block Weights

The largest block approach [16] picks the largest leaf block node that has the most number of words. N-EXT keeps text content of each leaf block node. At this stage the leaf block node with the most number of words can be selected as the news node. Although this choice is usually correct, it fails when another block, which contains other textual items (e.g., there can be several reader comments), contain more number of words than the actual news block. To address this problem, we assign a weight to each block and the one with the highest weight is selected as the news block.

We calculate block weights by paying attention to the block size and block similarity to the news title extracted from the RSS feed. - The use of similarity in such cases has a basis in the well-known vector space model [57].- Although

similarity of blocks to news description is more decisive (news descriptions contain more information about news contents than that of news titles), but we still go with the news titles since most RSS feeds do not contain news descriptions. We calculate the block weight of block i (w_i) using the formula (4.1).

$$w_i = \left(\beta \times \frac{s_i}{\max_{i\ \epsilon\ \{1,\ldots,n\}}(s_i)}\right) + \left((1-\beta) \times \frac{sim_i}{\max_{i\ \epsilon\ \{1,\ldots,n\}}(sim_i)}\right) \qquad (4.1)$$

In the formula given above, s_i is the size of block i in terms of number of words, sim_i is the similarity value of the block to the news title extracted from the RSS feed, n is the number of blocks in the web page, and β is the *"block weight assignment coefficient"* that controls the effect of size and similarity on the weight assigned to the block. w_i and β have a value between 0 and 1. We derive block weight by first normalizing block size and similarity of block to news title, and then assign weights to the normalized size and similarity values.

The similarity value calculation between blocks and news title is illustrated in Figure 4.3. In this figure, we assume that there are two candidate blocks, Block 1 and Block 2, where only one of them will be selected as the news block. In the same figure, *"a, b, c, d, e"* indicate the terms (stemmed words) that appear in the news title and blocks (more information on similarity value calculation is provided in the next section).

Before assigning a term frequency to each word, N-EXT first eliminates stopwords, which are the most frequent words of a language and are not meaningful alone but used for semantic integrity of sentences. Since these words exist frequently in the sentences, they affect the term frequency assignment in an unrealistic way. Thus, N-EXT eliminates stopwords from all leaf block nodes. (In the experiments, we use the union of two stopwords lists for Turkish [5] listed in Table A.1, and the Snowball [58] stopwords list for English listed in Table A.2.)

We use stemming in order to eliminate morphological variations of words and to obtain terms. We use the Zemberek [59] and Porter [60] stemmer for Turkish and English, respectively. Term frequency (actually relative term frequency) of

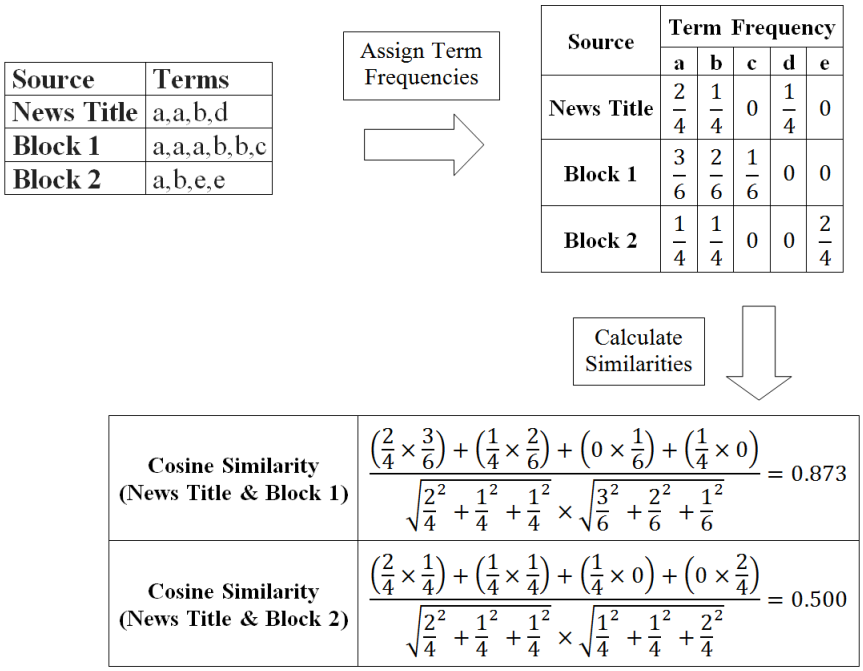


Figure 4.3: Example similarity calculation between candidate news blocks and news title.

each term in blocks is calculated by using the formula n_a/n_B, where n_a and n_B are, respectively, the frequency of the term and the total number of terms in the block (we use a similar approach for the news title and sentences when needed).

4.1.5 Extracting Content of the News Block

In this step N-EXT tries to detect the news content related information in the news block: the news block of a news web page may contain additional textual information not related to the news content (such as advertisements). In this context, N-EXT calculates the similarity value of the news block sentences to the news block itself.

$$CosineSimilarity(S,B) = \frac{\sum_{k=1}^{n}\left(tf_{s,k} \times tf_{b,k}\right)}{\sqrt{\sum_{k=1}^{n} tf_{s,k}^2 \times \sum_{k=1}^{n} tf_{b,k}^2}} \tag{4.2}$$

$$DiceSimilarity(S,B) = \frac{2 \times \sum_{k=1}^{n}\left(tf_{s,k} \times tf_{b,k}\right)}{\sum_{k=1}^{n} tf_{s,k}^2 + \sum_{k=1}^{n} tf_{b,k}^2} \tag{4.3}$$

$$JaccardSimilarity(S,B) = \frac{\sum_{k=1}^{n}\left(tf_{s,k} \times tf_{b,k}\right)}{\sum_{k=1}^{n} tf_{s,k}^2 + \sum_{k=1}^{n} tf_{b,k}^2 - \sum_{k=1}^{n}\left(tf_{s,k} \times tf_{b,k}\right)} \tag{4.4}$$

$$OverlapSimilarity(S,B) = \frac{\sum_{k=1}^{n}\left(tf_{s,k} \times tf_{b,k}\right)}{min\{\sum_{k=1}^{n} tf_{s,k}^2, \sum_{k=1}^{n} tf_{b,k}^2\}} \tag{4.5}$$

In the formulas (4.2), (4.3), (4.4), and (4.5), k represents the current term, n is the total number of terms in the news block, and tf is the term frequency assigned to term k. Notations S and B are both vectors representing a sentence and a news block, respectively. We treat each sentence in the news block as a query, and the news block itself as a document, and use the similarity measures listed in the formulas given above to calculate the similarity of each query to the document. The similarity between a query and a document represents the similarity of a sentence to the news block. An example for representing a document and its sentences as vectors, and calculating similarities between them by using each of four similarity measures is given in Figures B.1, B.2, B.3, B.4, and B.5.

After calculating the similarity value of each sentence to the news block, N-EXT compares similarity values calculated with a threshold value t, which is

calculated dynamically by taking the harmonic mean of similarity values of all sentences in the news block. If the similarity value of a sentence is less than t, then that sentence is treated as a noise, and eliminated from the news block. Note that we use harmonic mean to calculate the threshold value that is used to determine the relatedness of a sentence to the news content. The harmonic mean gives a similar weight to each data in the set. It shows the central point of all data in the set, and each data has an similar impact on the determination of the central point, not relative to its value, so that an outlier affects the central point like an ordinary data.

4.1.6 Detecting More Content in Other Blocks

N-EXT analyzes other leaf block nodes in addition to the selected news block to detect additional news content related sentences. Note that some of the news related elements, such as news description, may be placed in another leaf block node. N-EXT analyzes the contents of these blocks sentence by sentence. Each sentence is treated as a query and term frequencies are obtained sentence by sentence. If the similarity value of a sentence is greater than the threshold value detected at the previous step, that sentence is added to the extracted part. After analyzing all other leaf block nodes sentence by sentence, N-EXT finishes the news content extraction process.

4.2 Pseudocode of N-EXT

Pseudocode of N-EXT is given in Algorithm 1.

Algorithm 1 N-EXT Algorithm

```
loop
  Update RSS feeds by re-downloading them.
  Parse RSS feeds to obtain n1 titles and URL links of news web pages.
  Download n_1 HTML pages from URL links obtained.
  for HTML Page No = 1 to n_1 do
    Parse HTML page into a DOM tree.
    Traverse DOM tree to detect n_2 leaf block nodes.
    for Leaf Block Node No = 1 to n_2 do
      Extract text from the leaf block node.
      Eliminate noises from the extracted text.
      Assign a weight to the block by considering textual size and similarity
      to the news title of the cleaned text.
    end for
    Select the block with highest weight as the news block.
    Extract news content related sentences from the news block.
    Examine blocks other than the news block to detect rest of the news
    content if any exists.
  end for
end loop
```

Chapter 5

Evaluation Measures

5.1 NBD Evaluation Measures

We evaluate the news block detection performance of N-EXT by *NBD Accuracy*, which is the ratio of the number of true matches between manually labeled blocks and the blocks detected by N-EXT to the number of all labeled blocks. As an example, we have 100 sample web pages, and news blocks of all these 100 web pages are manually labeled by us. Then, N-EXT performs NBD on these sample web pages, and extracts the blocks detected as the news block from these web pages. Then, we check how many of the extracted blocks are labeled, i. e., how many blocks match with true news blocks. For instance, if 72 of 100 extracted blocks are labeled, then NBD accuracy = 72/100. To sum up, NBD accuracy is the ratio of total number of matched blocks to the number of all labeled blocks, as given in Formula (5.1).

$$NBDAccuracy = \left(\frac{n_{matched\ blocks}}{n_{all\ labeled\ blocks}} \right) \tag{5.1}$$

5.2 NCE Evaluation Measures

News contents extracted from news web pages are compared to the contents of the same web pages in the ground truth dataset to evaluate the NCE (news content extraction) performance of N-EXT. Figure 5.1 illustrates the terms used during comparisons.

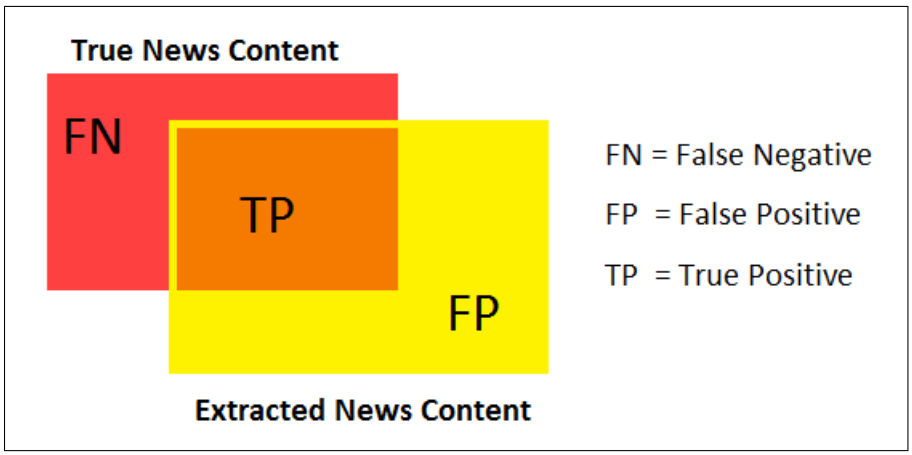


Figure 5.1: Illustration of the terms used in the set-based measures.

In this figure, *TP (True Positive)* is the set of relevant words (tokens, a relevant word is any word that appears in the ground truth version of the page) extracted from web page; *FP (False Positive)* is the set of unrelevant words extracted from web page; and *FN (false negative)* is the set of relevant words that could not be extracted from web page. Additionally, terms FN and TP together represents the true news content of a news web page, which is the set of all relevant words, and FP and TP together represents the news content extracted by N-EXT from new web page, which is the set of all extracted words. These terms are used in the set-based measures: precision, recall, and the F-measure as defined in the formulas (5.2), (5.3), and (5.4), respectively. In the formulas below, $|TP|$, $|FP|$, and $|FN|$ represent the word counts in the sets. Measures given in the formulas have values between 0 and 1, and 1 represents the best case [61]. A demonstration for calculation of these set-based measures is given in Figure B.8

$$Precision = \left(\frac{|TP|}{|TP| + |FP|} \right) \tag{5.2}$$

$$Recall = \left(\frac{|TP|}{|TP| + |FN|} \right) \tag{5.3}$$

$$F - measure = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall} \right) \tag{5.4}$$

We use the F-measure value to evaluate the NCE performance of N-EXT.

5.3 Means Comparison Measures

To assess the impact of similarity measures on NBD accuracy, we perform a means comparison using a one-way *Analysis of Variance (ANOVA)* with a *Scheffé* comparison [62]. ANOVA tests whether one or more sample means are significantly different from each other. It is similar to the *t-test*, but they differ from each other, since more than 2 groups can be tested simultaneously in ANOVA, whereas only 2 groups can be tested in t-test. Formulas given below are used to calculate one-way ANOVA.

$$SS_{total} = \left(\sum {x_1}^2 + \sum {x_2}^2 + ... + \sum {x_r}^2 \right) - \frac{\left(\sum x_1 + \sum x_2 + ... + \sum x_r \right)}{N} \tag{5.5}$$

$$SS_{among} = \left[\frac{\left(\sum x_1 \right)^2}{n_1} + \frac{\left(\sum x_2 \right)^2}{n_2} + ... + \frac{\left(\sum x_r \right)^2}{n_r} \right] - \frac{\left(\sum x_1 + \sum x_2 + ... + \sum x_r \right)}{N} \tag{5.6}$$

$$SS_{within} = SS_{total} - SS_{among} \tag{5.7}$$

$$df_{among} = r - 1 \tag{5.8}$$

$$df_{within} = N - r \tag{5.9}$$

$$MS_{among} = \frac{SS_{among}}{df_{among}} \tag{5.10}$$

$$MS_{within} = \frac{SS_{within}}{df_{within}} \tag{5.11}$$

$$F = \frac{MS_{among}}{MS_{within}} \tag{5.12}$$

In the formulas given above, SS represents *Sum of Squares* value, MS represents *Mean Square* value, df represents *Degrees of Freedom*, x represents an individual observation, r is the number of groups, N is total number of observations

in all groups, and n is the number of observations in a group. A demonstration for calculation of F score is given in Figure B.6. After calculating F score, it is compared to the value given in F table for $alpha = .05$ [63]. If calculated F score is bigger than the F score given in that table, then calculated F score is statistically significant. If calculated F score is statistically significant, it only indicates that at least two means are significantly different from each other, but it can't be known which mean pairs are significantly different from each other until a post-hoc test.

If a significant difference is found among sample means, *post hoc testing* is performed to determine which or how many sample means are different from each other. There are some post hoc testing procedures such as *Bonferroni test*, *Duncan's test*, *Tukey's HSD test*, and *Scheffé's test*. Scheffé's test is one of the most popular of the post hoc tests. Scheffé's test is generally used with unequal sample sizes, although it can be used with samples with equal sizes. Formula given below is used to calculate Scheffé's test.

$$F_{critical} = (k-1) \times F_{table} \tag{5.13}$$

$$F_{i,j} = \frac{(M_i - M_j)^2}{MS_{within} \times \left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{5.14}$$

In the formulas given above, k is the number of means, F_{table} is the value given in F table for $alpha = .05$ [63], n_i and n_j repsesent the sizes of samples i and j, respectively; and M_i and M_j represent the mean values of samples i and j, respectively. After calculating F scores for each pair of samples, each calculated $F_{i,j}$ score is compared to calculated $F_{critical}$ value. Then, for example, if only one of the calculated F scores is bigger than $F_{critical}$ value, such that $F_{1,3}$, then the means comparison between samples 1 and 3 is significantly different, but not the other comparisons. A demonstration for calculation of Scheffé's test is given in Figure B.7

Chapter 6

Experimental Environment and Experimental Results

In this section, we present the experimental environment and experimental results.

6.1 Experimental Environment

6.1.1 Implementation

We implemented N-EXT in Java language using the Eclipse IDE Helios Service Release 2^{nd} version, and performed our experiments on a computer which has an Intel Core 2 Quad Q9550@2.83 GHz CPU with 8 GB of main memory on Windows 7 64-bit operating system.

6.1.2 Test Collections

We perform our experiments on a ground truth test collection that we created during the course of this study. It consists of four components and they are manually

- extracted textual content of 3,500 Turkish news web pages published in 2012 (TR-Text),
- labeled blocks of 1,000 Turkish news web pages randomly chosen from TR-Text dataset (TR-Block),
- extracted textual content of 100 English news web pages published in 2012 (ENG-Text),
- labeled blocks of ENG-Text dataset (ENG-Block).

News web pages of the TR-Text and TR-Block are gathered from seven popular Turkish news websites which regularly use RSS feeds to disseminate the latest news:

- *CNN Türk* (http://www.cnnturk.com)
- *Milliyet* (http://www.milliyet.com.tr)
- *Sabah* (http://www.sabah.com.tr)
- *Samanyolu* (http://www.samanyoluhaber.com)
- *Star* (http://www.stargazete.com)
- *Yeni Şafak* (http://yenisafak.com.tr)
- *Zaman* (http://www.zaman.com.tr)

We try to gather different kinds of news documents by choosing news documents from different news categories into the ground truth dataset to observe the extraction performance of N-EXT over a wide variety of news documents. Since, textual size of the news contents differs from news category to news category; for example, economy, politics, and sport news web pages have larger textual sizes than those of magazine and technology news web pages. The distribution of the news documents among the news categories is listed in Table 6.1.

The news web pages of the ENG-Text and ENG-Block are from various news categories and are obtained from five popular world-wide English news websites:

- *BBC News* (http://www.bbc.com/news)

Category	News Websites						
	CN	**ML**	**SB**	**SM**	**ST**	**YŞ**	**ZM**
Agenda	0	0	80	80	80	80	80
Economy	70	80	80	70	80	60	75
Life	30	25	40	60	0	60	80
Local	80	80	0	0	0	0	0
Magazine	65	80	60	40	80	40	45
Politics	0	75	0	70	70	60	80
Sport	95	80	80	80	80	80	90
Technology	80	0	90	20	30	40	25
World	80	80	80	80	80	80	85
Total	**500**	**500**	**500**	**500**	**500**	**500**	**500**

Table 6.1: Distribution of news web pages to the news categories (CN=CNN Türk, ML=Milliyet, SB=Sabah, SM=Samanyolu, ST=Star, YŞ=Yeni Şafak, ZM=Zaman).

- *CNN* (http://www.cnn.com)
- *Fox News* (http://www.foxnews.com)
- *Los Angeles Times* (http://www.latimes.com)
- *The New York Times* (http://www.nytimes.com)

The size of the English dataset is smaller since it is our secondary test collection; however, its use in our work is important since it enables us to demonstrate that

- N-EXT is a language-independent NCE method,
- observations we had in one language are also applicable to another languages.

6.2 Experimental Results

Our experiments have two components: we first show that we successfully detect the news block, and after that we show our success in news content extraction.

6.2.1 News Block Detection (NBD) Results

We firstly evaluate the NBD performance of N-EXT and compare it with our baseline, the largest block approach [16], and show that it outperforms the baseline and is highly accurate. In the experiments, we use the TR-Block and ENG-Block datasets. In the TR-Block experiments, we use the k-fold cross-validation approach [64] for choosing the block weight assignment coefficient (β), which allocates importance to block size and similarity in the calculation of block weights. In k-folding approach, we use 10 for k. During the experiments, TR-Block dataset is partitioned into ten subsets, each having equal number of news web pages (1000/10 = 100). For each of the ten experiments, nine of the subsets are used for training, and one of the subsets is used for testing. We repeat these experiments for each one of the four similarity measures used in the calculation of block weights. Figure 6.1 demonstrates the k-fold cross validation approach.

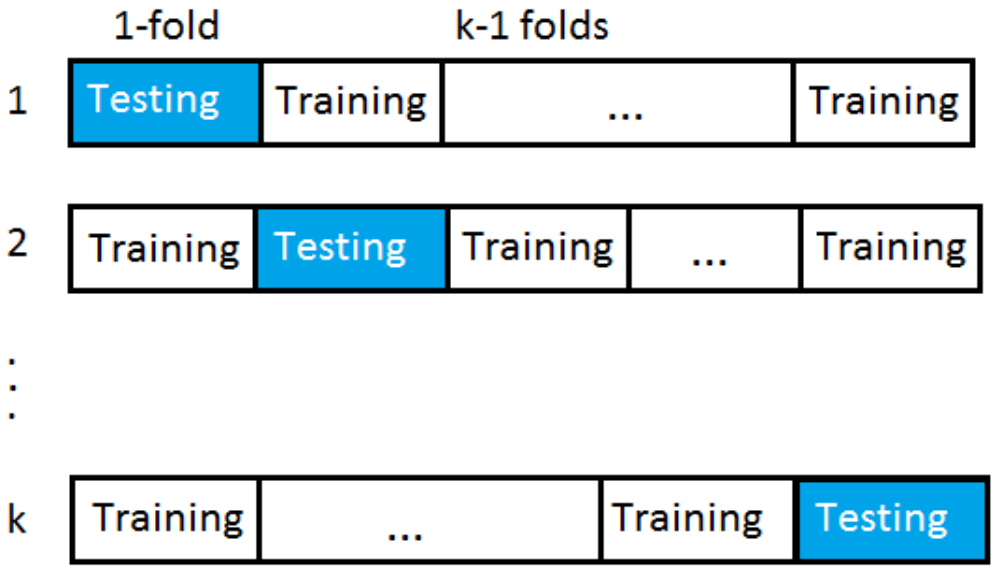


Figure 6.1: K-fold cross-validation approach.

Table 6.2 gives the detailed NBD results obtained using TR-Block dataset for different β values from 0.0 to 1.0 in the training and testing using the Dice similarity measure in the calculation of block weights (in test results we give two digits after the decimal point since in each fold we have 100 test cases). Besides,

β / k	β Values										
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
1	0.922	0.928	0.936	0.940	0.949	0.955	**0.963**	0.951	0.942	0.929	0.918
2	0.919	0.926	0.937	0.941	0.949	0.958	**0.967**	0.950	0.949	0.924	0.912
3	0.925	0.929	0.939	0.941	0.952	0.954	0.955	**0.957**	0.948	0.932	0.921
4	0.923	0.926	0.937	0.943	0.950	0.954	**0.961**	0.952	0.947	0.930	0.922
5	0.921	0.925	0.934	0.938	0.947	**0.955**	0.954	0.952	0.946	0.931	0.924
6	0.923	0.932	0.939	0.946	0.954	0.958	**0.969**	0.959	0.948	0.938	0.928
7	0.924	0.927	0.935	0.943	0.949	0.955	**0.963**	0.951	0.942	0.929	0.918
8	0.924	0.931	0.938	0.942	0.944	0.951	0.953	**0.956**	0.947	0.933	0.925
9	0.923	0.926	0.938	0.945	0.949	0.952	0.954	**0.955**	0.946	0.939	0.931
10	0.925	0.929	0.936	0.945	0.947	0.956	**0.968**	0.953	0.944	0.927	0.920
Avg.	0.923	0.928	0.937	0.942	0.949	0.955	**0.961**	0.954	0.946	0.931	0.922

a) NBD accuracy training results.

β / k	β Values										
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
1	0.90	0.91	0.92	0.93	0.94	0.95	**0.96**	0.95	0.93	0.93	0.92
2	0.91	0.92	0.94	0.94	0.95	0.96	**0.97**	0.95	0.94	0.92	0.91
3	0.92	0.93	0.94	0.94	0.94	0.95	0.95	**0.96**	0.95	0.93	0.92
4	0.90	0.92	0.93	0.94	0.94	0.94	**0.96**	0.95	0.94	0.93	0.92
5	0.91	0.93	0.93	0.94	0.94	**0.95**	**0.95**	**0.95**	0.93	0.93	0.92
6	0.91	0.93	0.94	0.95	0.95	0.96	**0.97**	0.96	0.95	0.94	0.93
7	0.92	0.92	0.93	0.94	0.95	0.95	**0.96**	0.95	0.94	0.93	0.91
8	0.90	0.91	0.93	0.94	0.94	0.95	**0.96**	0.95	0.94	0.92	0.92
9	0.91	0.92	0.93	0.94	**0.95**	**0.95**	**0.95**	**0.95**	0.94	0.93	0.92
10	0.92	0.93	0.93	0.94	0.94	0.95	**0.96**	0.95	0.94	0.92	0.91
Avg.	0.910	0.922	0.932	0.940	0.944	0.951	**0.959**	0.952	0.940	0.928	0.918

b) NBD accuracy testing results.

Table 6.2: News block detection (NBD) accuracy training and testing results of N-EXT with TR-Block dataset (without hyperlink texts) using Dice similarity measure and 10-fold cross-validation.

Tables C.1, C.2, and C.3 give detailed NBD results obtained on TR-Block dataset using other similarity measures (Cosine, Dice, and Jaccard) with the same parameters.

As it is seen in Table 6.2, in most cases, we obtained the best NBD accuracy when $\beta = 0.6$ in training and testing. Besides, the other similarity measures also give the best NBD accuracy in most cases when $\beta = 0.6$. For brevity we only show the detailed results for the Dice similarity measures.

Table 6.3 shows the NBD accuracy test results using the TR-Block dataset

k	Similarity Measures			
	Cosine	**Dice**	**Jaccard**	**Overlap**
1	0.93	**0.96**	**0.96**	0.91
2	0.93	**0.97**	0.96	0.91
3	0.92	**0.95**	**0.95**	0.90
4	0.92	**0.96**	0.95	0.91
5	0.91	**0.95**	0.94	0.90
6	0.93	**0.97**	0.96	0.92
7	0.93	**0.96**	**0.96**	0.91
8	0.92	**0.95**	**0.95**	0.92
9	0.92	**0.95**	0.94	0.91
10	0.94	**0.97**	0.96	0.93
Average	0.925	**0.959**	0.953	0.911

Table 6.3: News block detection (NBD) accuracy testing results of N-EXT with TR-Block dataset (without hyperlink texts) using different similarity measures in the calculation of weight of a block when $\beta = 0.6$.

for all similarity measures with $\beta = 0.6$. The results show that we obtain the best NBD accuracy using the Dice similarity measure. Since we are comparing the NBD means (average values) of four similarity measures, we should do a means comparison using a one way *ANOVA* with a *Scheffé* comparison [62]. The test shows that the means of NBD results obtained by using the Cosine, Dice, Jaccard, and Overlap measures are significantly different at $p < 0.05$; only exception is the Dice and Jaccard similarity measures, i.e., they do not have significantly different means. Although Dice and Jaccard are not statistically significantly different, we prefer Dice since it gives the highest average NBD accuracy in the tests.

Our NBD approach performs best when $\beta = 0.6$ as it is seen in Table 6.2. The value of β is the only coefficient that needs to be determined by training. Table 6.4 gives the detailed NBD results obtained using ENG-Block dataset for different β values from 0.0 to 1.0 in the training and testing using the Dice similarity measure in the calculation of block weights. Additionally, Tables C.4, C.5, and C.6 give detailed NBD results obtained on ENG-Block dataset using other similarity measures (Cosine, Dice, and Jaccard) with the same parameters. The test results given in Table 6.4, arguably show that N-EXT detects the news blocks most accurately when $\beta = 0.6$ as suggested by the results based on the

β / ht	β Values										
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
W. H. T.[1]	0.77	0.78	0.79	0.79	0.80	0.80	**0.82**	0.81	0.80	0.79	0.79
W/O. H. T.[1]	0.88	0.88	0.89	0.91	0.91	0.91	**0.92**	0.91	0.89	0.88	0.87

Table 6.4: News block detection (NBD) accuracy results of N-EXT with ENG-Block dataset using Dice similarity measure.

Datasets		**Approaches**		
		Only Similarity ($\beta = 0$)	**Size & Similarity** ($\beta = 0.6$)	**Only Size (Baseline)** ($\beta = 1$)
Turkish	with hyperlink texts	0.818	0.856	0.809
	w/o hyperlink texts	0.910	**0.959**	0.918
English	with hyperlink texts	0.770	0.820	0.790
	w/o hyperlink texts	0.880	**0.920**	0.870

Table 6.5: Summary of the news block detection (NBD) accuracy results of N-EXT with TR-Block and ENG-Block datasets using Dice similarity measure.

TR-Block dataset.

Summary of the NBD results obtained in the experiments is provided in Table 6.5. The experimental results show that our NBD approach outperforms the baseline in both languages.

6.2.1.1 Additional Observations Based on NBD Experiments

It is observed that our approach have difficulty in detecting the news block when the textual size of news is comparatively smaller than the other elements of web pages, especially hyperlink texts. This deteriorates the NBD effectiveness. On the other hand, hyperlink text percentage of news contents is less than one percent of the entire text we extract from news. Hence deleting hyperlink texts is a negligible loss in content extraction. The experimental results given in Tables 6.4 and 6.5 show that cleaning hyperlink texts improves the NBD performance of N-EXT. Accordingly the NCE experiments presented in the next subsection are also performed after deleting hyperlinks.

The experimental results given in Table 6.5 also demonstrate that NBD with the Turkish dataset has a higher accuracy than that of the English dataset. After

[1]W. H. T = With Hyperlink Texts, W/O. H. T = Without Hyperlink Texts

examining both Turkish and English datasets, we conclude that the reason of the difference in the detection accuracy is due to higher heterogeneity of English news web pages. Because of that, the detection of the news block is negatively affected; however, note that the decrease is small.

6.2.2 News Content Extraction (NCE) Results

Average F-measure values obtained in the NCE experiments performed on the TR-Text dataset is given in Table 6.6. We observe the highest average F-measure values with the Dice similarity measure.

Datasets	**Stemming**	**Similarity Measures**			
		Cosine	**Dice**	**Jaccard**	**Overlap**
Turkish	**With Stemming**	0.908	**0.922**	0.917	0.902
	Without Stemming	0.897	0.914	0.910	0.894
English	**With Stemming**	0.886	**0.907**	0.902	0.874
	Without Stemming	0.880	0.899	0.893	0.868

Table 6.6: Average F-measure values for news content extraction (NCE) using TR-Text and ENG-Text datasets.

Moreover, when we perform the NCE experiments on the ENG-Text dataset, we obtain slightly better performance with stemming. Like that of the TR-Text dataset, we again obtain the best performance with the Dice similarity measure in the ENG-Text dataset.

In the NCE experiments, we observe that the experiments done on the Turkish dataset obtain higher F-measure values than those of the English dataset. This is again due to higher heterogeneity of the English news web pages: the heterogeneity of pages complicates the extraction.

Additionally, Table 6.7 details the average F-measure values obtained with using stemming in the experiments performed on the TR-Text dataset. These results show that NCE accuracy shows variation among news websites. Average values in the table demonstrate that NCE is slightly more accurate for the news web pages of Star, Yeni Şafak, and Zaman since these websites store all news content elements only in one block; however, in the other websites news content

News Websites	Similarity Measures				Average
	Cosine	Dice	Jaccard	Overlap	
CNN Türk	0.902	0.927	0.919	0.895	0.911
Milliyet	0.901	0.909	0.907	0.897	0.904
Sabah	0.901	0.912	0.910	0.899	0.906
Samanyolu	0.907	0.915	0.912	0.902	0.909
Star	0.918	0.938	0.926	0.908	**0.923**
Yeni Şafak	0.912	0.929	0.924	0.905	0.918
Zaman	0.915	0.923	0.921	0.909	0.917
Average	0.908	**0.922**	0.917	0.902	0.913

Table 6.7: Average F-measure values for different news websites obtained with using stemming.

elements are distributed among more than one block. Average F-measure values obtained without using stemming are given in C.10

6.2.3 Multithreading Results

To analyze and evaluate the impact of multithreading on total extraction time of N-EXT, we also implemented the multithreaded version of the stages except the first stage of the news extraction process. We prefer using *Single Instruction Stream, Multiple Data Stream (SIMD)* architecture [65], which has a single control unit that dispatches the same instruction to various processors (that work on different data) demonstrated in Figure 6.2.

As seen from the model given in Figure 6.2, a control unit dispatches the same instruction, which is extracting the news content from news web pages in our example, to all processors including the one on which that control unit executes. Then, each processor works on its own set of news web pages, and extracts contents from them.

We prepared an additional dataset for multithread experiments by randomly chosing 100 news pages from TR-Text dataset. In multithread experiments, we compute total extraction time of extracting news contents from this additional dataset for each thread count given to the multhreaded implementation of N-EXT as a parameter. Tests for each thread count parameter are repeated for 10 times.

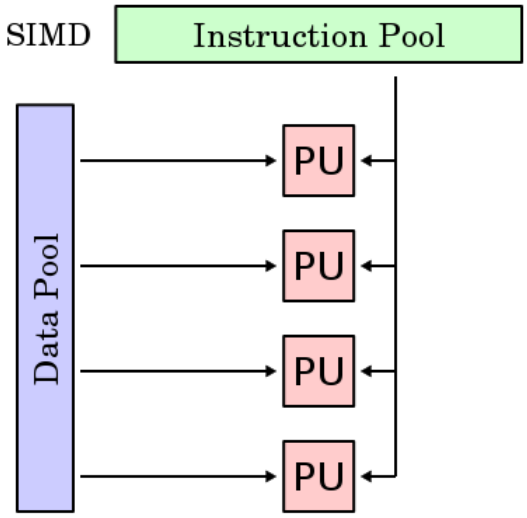


Figure 6.2: A typical SIMD architecture.

Figure 6.3 shows the results of multithreading experiments, which are the mean values of results obtained in those 10 experiments.

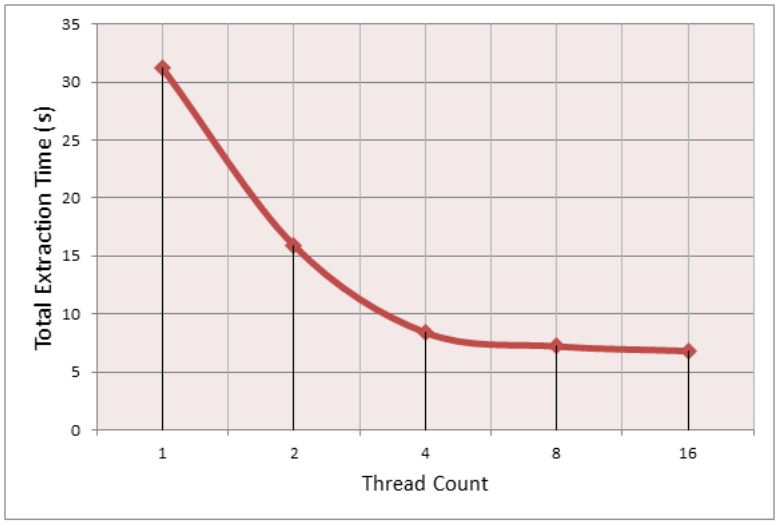


Figure 6.3: Total extraction time VS. thread count.

As it is seen in Figure 6.3, total extraction time decreases obvioulsy until selecting 4 as the thread count for the extraction process. Then, total extraction time becomes nearly stable. After performing a one-way ANOVA with Scheffe's comparison to the mean values obtained for different thread counts, we observed that

- mean values obtained for thread count = 1 **are significantly different** than mean values obtaine for thread count = 2,
- mean values obtained for thread count = 2 **are significantly different** than mean values obtaine for thread count = 4,
- mean values obtained for thread count = 4 **are not significantly different** than mean values obtaine for thread count = 8,
- mean values obtained for thread count = 8 **are not significantly different** than mean values obtaine for thread count = 16.

As the comparison results show, we cannot gain any significant decrease in total extraction time after selecting the thread count as 4. Main reason of this is that the computer we are running our experiments have a CPU with 4 processors. When we select thread count parameter as 4, the processor assigned as the dispatcher, dispatches a single thread to each of four processors. After analyzing total load of an extraction process, we observed that N-EXT reserves approximately %90-95 of processor load during the extraction process. That much load ratio does not cause any bottleneck for processor load. But, this condition changes when we choose to divide the extraction process into 8 threads. This time, dispatcher dispatches 2 threads to each of four processors. When more than one thread share a processor, load of that processor is divided into number of threads using that processor. However, as we observed, N-EXT needs more than %90 of a processor load to work properly. Hence, a bottleneck for the processor load occurs, which causes a latency in execution of instructions dispatched to each processor. As a result, although total number of news pages being executed is increased, total extraction time of each news pages also increases due to latency occuring. Therefore, we could not gain any significant decrease in total extraction time for more than 4 threads. As a result, we select 4 as the thread count parameter for the multithreaded implementation of N-EXT.

Chapter 7

Bilkent News Portal

7.1 Configuration of Bilkent News Portal

As it is mentioned before, news content extraction (NCE) is used in our news portal, called *Bilkent News Portal* [10], which uses RSS feeds to gather news web pages from various different news websites, extracts news contents from these news web pages, and displays the contents to the web users.

Configuration of Bilkent News Portal is demonstrated in Figure 7.1. As it is seen in Figure 7.1, Bilkent News Portal has three main PCs: *a dispatcher*, *a PHP server*, and *a database server*. All of these three PCs have Linux operating system installed on themselves. Dispatcher is the only PC of Bilkent News Portal that can be directly accessed over Internet. *Secure Shell (SSH)* [66] is used to make a remote access to a Linux machine, since we use a free Telnet/SSH Client tool, called *PuTTY* [67] to access the dispatcher. After accessing the dispatcher, "*ssh host_address*" Linux command is used to make remote accesses to PHP and database servers. IP configurations of these three PCS are listed on Table 7.1.

When a user opens one of the web browsers, and requests a connection to Bilkent News Portal using "*http://139.179.21.201/PortalTest/*" address; the dispatcher directs the browser request to the PC on which PHP server is installed. Then, PHP server PC displays Bilkent News Portal web pages to the user.

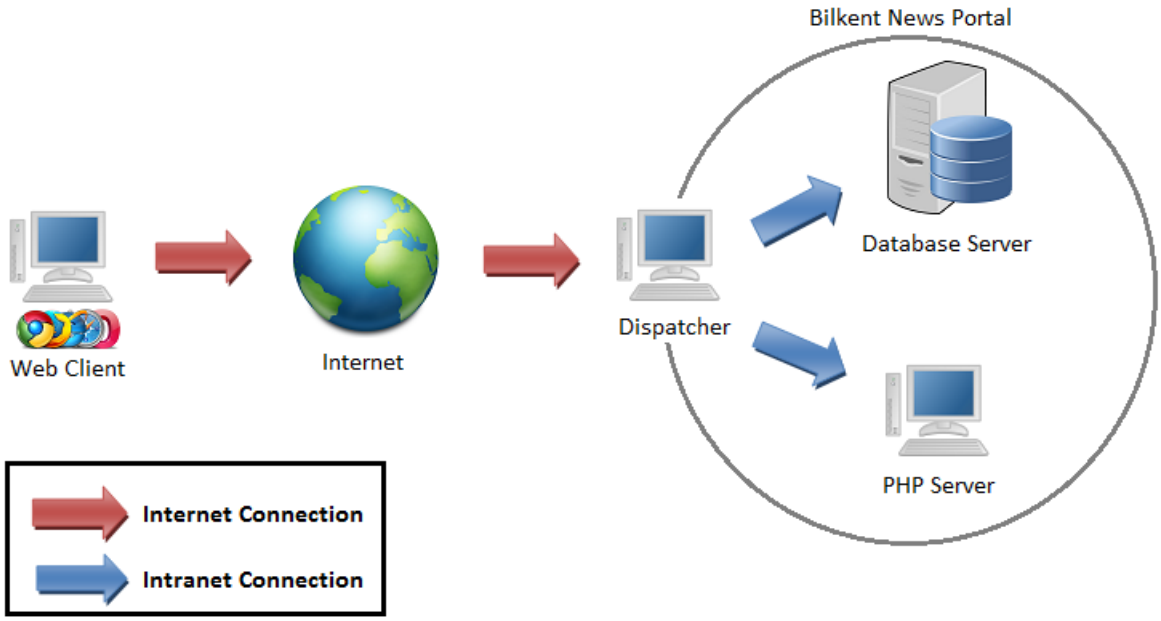


Figure 7.1: Configuration of Bilkent News Portal.

Definition	**PC name**	**Host Address**
Dispatcher	portal-alive	139.179.21.201
PHP Server	news-portal1	139.168.20.100
Database Server	news-portal2	139.168.20.101

Table 7.1: PC list of Bilkent News Portal.

Java programs, which perform basic operations such as news content extraction, topic tracking, novelty detection, etc., runs on the PC on which database server is installed. The program proposed in this thesis, N-EXT, is the Java program, which performs news content extraction. Firstly, it downloads HTML web pages to the same PC on which it runs. Then, it extracts news contents from those web pages, and inserts them into the news database. News contents stored in that database is displayed to the user via PHP server.

7.2 Deployment of N-EXT to the Portal

To deploy N-EXT to Bilkent News Portal, following steps are required:

- Use *PuTTY* [67] to make a remote access to "*139.179.21.201*", which is the IP address of Bilkent News Portal's dispatcher PC,
- Type "root" for login name and "**********" for password, and press "ENTER" button,
- Use "ssh" linux command to make another remote access from accessed dispatcher (*ssh 192.168.20.101*),
- Type "**********" for password, and press "ENTER" button,
- Change current directory to the directory that contain news extraction source files (*cd /var/www/PortalTest/workspaceCrawlerParse/RSSCrawlerParser/src/*),
- Put related Java files (*Parser.java*) into this directory,
- Change current directory to another directory (*cd /var/www/PortalTest/*),
- Open "NewsPortal.sh" shell script file by using a text editor, such as "vi" (*vi NewsPortal.sh*),
- Edit the following command, which executes current Java file (Download_News.java) that performs news extraction with respect to the changes made in that file (java -cp ... Download_News 4),
- To save the changes made, first press "ESC" button to enter into command mode, and then type ":w" and press "ENTER" button,
- Finally, to quit from the editor, first press "ESC" button to enter into command mode, and then type ":q" and press "ENTER" button.
- If you want to quit without saving, first press "ESC" button to enter into command mode, and then type ":q!" and press "ENTER" button.
- To test the portal, connect to the IP address "*http://139.179.21.201/PortalTest*", which is the address of portal's main page, from a browser.

Chapter 8

Conclusion

Content extraction accuracy of news web pages is important since it directly affects the performance of information retrieval and web mining modules of news aggregators. In this thesis, we propose a template-independent content extraction method (N-EXT) for news web pages. Our approach avoids the major problems of template-based extraction methods, such as human intervention and regular maintenance. Our method N-EXT examines all web page blocks to detect the news block that contains the major part of the news content. For this purpose, we assign weights to blocks using their size and similarity to news title. The block with the maximum weight is selected as the news block. For quantifying the importance of these two weight components and selecting a similarity measure we use the k-fold cross validation approach and one way ANOVA with a Scheffé comparison, respectively. We show that removing hyperlink texts and stemming respectively improves the NBD (news block detection) and NCE (news content extraction) accuracy. Besides, we also show that multithreading positively effects total extraction time up to 4 threads.

We experimentally demonstrate the effectiveness of N-EXT on pages obtained from several Turkish and English news websites. The experimental results show that our method is robust and highly accurate and can be used in real life applications. In this study, we also provide an NCE test collection that we will share with other researchers.

In future work, our approach can be modified according to the needs of other web information aggregators such as blog portals [68]. The extraction accuracy of N-EXT may be further increased by using other similarity measures such as earth movers distance (EMD) measure [69], or combining various measures together to calculate similarity of sentences to the news block.

Bibliography

[1] A. Aqarwal, "The size of internet to double every 5 years." http://www.labnol.org/internet/internet-size-to-double-every-5-years/6569/, 2009.

[2] P. R. Center, "Online papers modestly boost newspaper readership." http://people-press.org/2006/07/30/online-papers-modestly-boost-newspaper-readership/, 2006.

[3] T. Economist, "Bulletins from the future." http://www.economist.com/node/18904136/, 2011.

[4] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A survey of web information extraction systems," *IEEE Trans. on Knowl. and Data Eng.*, vol. 18, no. 10, pp. 1411–1428, 2006.

[5] F. Can, S. Kocberber, E. Balcik, C. Kaynak, H. C. Ocalan, and O. M. Vursavas, "Information retrieval on turkish texts," *J. Am. Soc. Inf. Sci. Technol.*, vol. 59, no. 3, pp. 407–421, 2008.

[6] F. Can, S. Kocberber, O. Baglioglu, S. Kardas, H. C. Ocalan, and E. Uyar, "New event detection and topic tracking in turkish," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 4, pp. 802–819, 2010.

[7] C. Aksoy, F. Can, and S. Kocberber, "Novelty detection for topic tracking," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 4, pp. 777–795, 2012.

[8] G. Ercan and F. Can, "Cover coefficient-based multi-document summarization," in *Proceedings of the 31th European Conference on IR Research on*

Advances in Information Retrieval, ECIR '09, (Berlin, Heidelberg), pp. 670–674, Springer-Verlag, 2009.

[9] E. Varol, F. Can, C. Aykanat, and O. Kaya, "Codet: sentence-based containment detection in news corpora," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, (New York, NY, USA), pp. 2049–2052, ACM, 2011.

[10] F. Can, S. Kocberber, O. Baglioglu, S. Kardas, H. C. Ocalan, and E. Uyar, "Bilkent news portal: a personalizable system with new event detection and tracking capabilities," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, (New York, NY, USA), pp. 885–885, ACM, 2008.

[11] H. Çağdaş Öcalan, "Bilkent news portal: A system with new event detection and tracking capabilities," Master's thesis, Bilkent University, 2009.

[12] B. Liu, *Web Data Mining*, ch. 9. New York, NY, USA: Springer, second ed., 2011.

[13] H. Han, T. Noro, and T. Tokuda, "An automatic web news article contents extraction system based on rss feeds," *J. Web Eng.*, vol. 8, no. 3, pp. 268–284, 2009.

[14] S. Vadrevu, S. Nagarajan, F. Gelgi, and H. Davulcu, "Automated metadata and instance extraction from news web sites," *Web Intelligence, IEEE / WIC / ACM International Conference on*, vol. 0, pp. 38–41, 2005.

[15] A. Arasu and H. Garcia-Molina, "Extracting structured data from web pages," in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, SIGMOD '03, (New York, NY, USA), pp. 337–348, ACM, 2003.

[16] L. Ziyi, S. Beijun, T. Xinhuai, and C. Delai, "Automatic web news extraction using blocking tag," in *Proceedings of the 2009 Second International Conference on Machine Vision*, ICMV '09, (Washington, DC, USA), pp. 74–78, IEEE Computer Society, 2009.

[17] A. H. F. Laender, B. A. Ribeiro-Neto, A. S. da Silva, and J. S. Teixeira, "A brief survey of web data extraction tools," *SIGMOD Rec.*, vol. 31, no. 2, pp. 84–93, 2002.

[18] G. O. Arocena and A. O. Mendelzon, "Weboql: restructuring documents, databases, and webs," *Theor. Pract. Object Syst.*, vol. 5, no. 3, pp. 127–141, 1999.

[19] V. Crescenzi and G. Mecca, "Grammars have exceptions," *Inf. Syst.*, vol. 23, no. 9, pp. 539–565, 1998.

[20] J. Hammer, J. McHugh, and H. Garcia-Molin, "Semistructured data: the tsimmis experience," in *Proceedings of the First East-European conference on Advances in Databases and Information systems*, ADBIS'97, (Swinton, UK, UK), pp. 22–22, British Computer Society, 1997.

[21] G. Huck, P. Fankhauser, K. Aberer, and E. J. Neuhold, "Jedi: Extracting and synthesizing information from the web," in *COOPIS '98: Proceedings of the 3rd IFCIS International Conference on Cooperative Information Systems*, (Washington, DC, USA), pp. 32–43, IEEE Computer Society, 1998.

[22] B. Ludäscher, R. Himmeröder, G. Lausen, W. May, and C. Schlepphorst, "Managing semistructured data with florid: a deductive object-oriented perspective," *Inf. Syst.*, vol. 23, no. 9, pp. 589–613, 1998.

[23] V. Crescenzi, G. Mecca, and P. Merialdo, "Roadrunner: Towards automatic data extraction from large web sites," in *Proceedings of the 27th International Conference on Very Large Data Bases*, VLDB '01, (San Francisco, CA, USA), pp. 109–118, Morgan Kaufmann Publishers Inc., 2001.

[24] A. Sahuguet and F. Azavant, "Building intelligent web applications using lightweight wrappers," *Data Knowl. Eng.*, vol. 36, no. 3, pp. 283–316, 2001.

[25] L. Liu, C. Pu, and W. Han, "Xwrap: An xml-enabled wrapper construction system for web information sources," in *Proceedings of the 16th International Conference on Data Engineering*, ICDE '00, (Washington, DC, USA), pp. 611–621, IEEE Computer Society, 2000.

[26] M. E. Califf and R. J. Mooney, "Relational learning of pattern-match rules for information extraction," in *Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*, AAAI '99/IAAI '99, (Menlo Park, CA, USA), pp. 328–334, American Association for Artificial Intelligence, 1999.

[27] D. Freitag, "Machine learning for information extraction in informal domains," *Mach. Learn.*, vol. 39, no. 2-3, pp. 169–202, 2000.

[28] S. Soderland, "Learning information extraction rules for semi-structured and free text," *Mach. Learn.*, vol. 34, no. 1-3, pp. 233–272, 1999.

[29] I. Muslea, S. Minton, and C. A. Knoblock, "Hierarchical wrapper induction for semistructured information sources," *Autonomous Agents and Multi-Agent Systems*, vol. 4, no. 1-2, pp. 93–114, 2001.

[30] N. Kushmerick, "Wrapper induction: efficiency and expressiveness," *Artif. Intell.*, vol. 118, no. 1-2, pp. 15–68, 2000.

[31] C.-N. Hsu and M.-T. Dung, "Generating finite-state transducers for semi-structured data extraction from the web," *Inf. Syst.*, vol. 23, no. 9, pp. 521–538, 1998.

[32] S. Zheng, R. Song, J.-R. Wen, and C. L. Giles, "Efficient record-level wrapper induction," in *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, (New York, NY, USA), pp. 47–56, ACM, 2009.

[33] B. Adelberg, "Nodose–a tool for semi-automatically extracting structured and semistructured data from text documents," in *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, SIGMOD '98, (New York, NY, USA), pp. 283–294, ACM, 1998.

[34] A. H. F. Laender, B. Ribeiro-Neto, and A. S. da Silva, "Debye - date extraction by example," *Data Knowl. Eng.*, vol. 40, no. 2, pp. 121–154, 2002.

[35] K. D. Hossam Ibrahim and A.-R. Madany, "Automatic extraction of textual elements from news web pages," in *Proceedings of the Sixth International*

Conference on Language Resources and Evaluation (LREC'08), (Marrakech, Morocco), European Language Resources Association (ELRA), may 2008.

[36] J. Gibson, B. Wellner, and S. Lubar, "Adaptive web-page content identification," in *Proceedings of the 9th annual ACM international workshop on Web information and data management*, WIDM '07, (New York, NY, USA), pp. 105–112, ACM, 2007.

[37] A. Spengler and P. Gallinari, "Learning to extract content from news webpages," in *Proceedings of the 2009 International Conference on Advanced Information Networking and Applications Workshops*, WAINA '09, (Washington, DC, USA), pp. 709–714, IEEE Computer Society, 2009.

[38] Y. Xue, Y. Hu, G. Xin, R. Song, S. Shi, Y. Cao, C.-Y. Lin, and H. Li, "Web page title extraction and its application," *Inf. Process. Manage.*, vol. 43, no. 5, pp. 1332–1347, 2007.

[39] A. Spengler, A. Bordes, and P. Gallinari, "A comparison of discriminative classifiers for web news content extraction," in *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, (Paris, France), pp. 172–175, Le Centre De Hautes Etudes Internationales D'Informatique Documantaire, 2010.

[40] J. Parapar and Àlvaro Barreiro, "An effective and efficient web news extraction technique for an operational NewsIR system," in *Proceedings of the XIII Conferencia de la Asociación Española para la Inteligencia Artificial CAEPIA*, vol. 2, (Salamanca, Spain), pp. 319–328, AEPIA, 2007.

[41] N. Gupta and S. Hilal, "A heuristic approach for web content extraction," *International Journal of Computer Applications*, vol. 15, no. 5, pp. 20–24, 2011.

[42] T. Gottron, "Combining content extraction heuristics: the combine system," in *Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services*, iiWAS '08, (New York, NY, USA), pp. 591–595, ACM, 2008.

[43] Q. Wu, X.-s. Chen, K. Zhu, and C.-h. Wang, "Relevance-based content extraction of html documents," *Journal of Central South University*, vol. 19, pp. 1921–1926, 2012.

[44] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[45] D. C. Reis, P. B. Golgher, A. S. Silva, and A. F. Laender, "Automatic web news extraction using tree edit distance," in *Proceedings of the 13th international conference on World Wide Web*, WWW '04, (New York, NY, USA), pp. 502–511, ACM, 2004.

[46] Q. Lan, "Extraction of news content for text mining based on edit distance," *Journal of Computational Information Systems*, vol. 6, no. 11, pp. 3761–3777, 2010.

[47] S. Zheng, R. Song, and J.-R. Wen, "Template-independent news extraction based on visual consistency," in *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*, AAAI'07, (Vancouver, British Columbia, Canada), pp. 1507–1512, AAAI Press, 2007.

[48] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, "Extracting content structure for web pages based on visual representation," in *Proceedings of the 5th Asia-Pacific web conference on Web technologies and applications*, APWeb'03, (Berlin, Heidelberg), pp. 406–417, Springer-Verlag, 2003.

[49] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proceedings of the Second European Conference on Computational Learning Theory*, EuroCOLT '95, (London, UK), pp. 23–37, Springer-Verlag, 1995.

[50] S. Debnath, P. Mitra, and C. L. Giles, "Automatic extraction of informative blocks from webpages," in *Proceedings of the 2005 ACM symposium on Applied computing*, SAC '05, (New York, NY, USA), pp. 1722–1726, ACM, 2005.

[51] S.-H. Lin and J.-M. Ho, "Discovering informative content blocks from web

documents," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, (New York, NY, USA), pp. 588–593, ACM, 2002.

[52] C.-N. Ziegler and M. Skubacz, "Content extraction from news pages using particle swarm optimization on linguistic and structural features.," in *Web Intelligence*, pp. 242–249, IEEE Computer Society, 2007.

[53] S. Shen and H. Zhang, "Block-level linkes based content extraction," *Parallel Architectures, Algorithms and Programming, International Symposium on*, vol. 0, pp. 330–333, 2011.

[54] W. W. W. Consortium, "Document object model (dom)." `http://www.w3.org/DOM/`, 2005.

[55] J. H. parser, "Jericho html parser." `http://jericho.htmlparser.net/docs/index.html`, 2011.

[56] P. E. Black, "Depth-first search." `http://www.nist.gov/dads/HTML/depthfirst.html`, 2010.

[57] G. Salton, *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989.

[58] M. Porter, "A stop word list." `http://snowball.tartarus.org/algorithms/english/stop.txt`, 2002.

[59] A. A. Akn and M. D. Akn, "Zemberek, an open source nlp framework for turkic languages." `http://zemberek.googlecode.com/files/zemberek_makale.pdf`, 2007.

[60] M. Porter, "The porter stemming algorithm." `http://tartarus.org/martin/PorterStemmer/`, 2007.

[61] C. J. van Rijsbergen, *Information Retrieval*, ch. 1. London, UK: Butterworths, second ed., 1979.

[62] B. L. Bowerman and R. T. O'Connell, *Linear Statistical Models: An Applied Approach*. Pacific Grove, CA, USA: Duxbury, second ed., 1990.

[63] StatSoft, "Compare distribution tables." http://www.statsoft.com/textbook/distribution-tables/, 2012.

[64] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[65] J. L. Hennessy and D. A. Patterson, *Thread-level paralellism*, ch. 5. Morgan Kaufmann, fifth ed., 2011.

[66] Wikipedia, "Secure shell." http://en.wikipedia.org/wiki/Secure_Shell#History_and_development, 2012.

[67] S. Tatham, "Putty: A free telnet/ssh client." http://www.chiark.greenend.org.uk/~sgtatham/putty/, 2011.

[68] K. Hofmann and W. Weerkamp, "Content extraction for information retrieval in blogs and intranets." http://ilps.science.uva.nl/sites/default/files/cikm2008-corpuscleaning_0.pdf, 2008. Technical report.

[69] X. Wan, "A novel document similarity measure based on earth mover's distance," *Inf. Sci.*, vol. 177, no. 18, pp. 3718–3730, 2007.

Appendix A

Data

A.1 Stopwords Lists

To eliminate stopwords from the sentences, two stopwords list is used: one for Turkish news, and another for English news. Turkish and English stopwords lists consist of 217 and 221 words, respectively, listed below.

acaba	birşey	demek	her	mu	oysa	şimdi
altı	birşeyi	diğer	herkes	mü	oysaki	şöyle
ama	biz	diğeri	herkese	nasıl	öbürü	şu
ancak	bize	diğerleri	herkesi	ne	ön	şuna
arada	bizi	diye	hiç	neden	önce	şunda
artık	bizim	dokuz	hiçbiri	nedir	ötürü	şundan
ayrıca	böyle	dolayı	hiçbirine	neler	öyle	şunlar
asla	böylece	dolayısıyla	hiçbirini	nerde	pek	şunu
aslında	böylesi	dört	için	nerede	peki	şunun
az	bu	eğer	içinde	nereden	rağmen	tabi
bana	budur	elbette	iki	nereye	sadece	tamam
bazen	buna	en	ile	nesi	sana	tarafından
bazı	bunda	fakat	ilgili	neyse	sanki	tüm
bazıları	bundan	falan	ise	niçin	sekiz	tümü
bazısı	bunlar	felan	işte	niye	sen	üç
belki	bunları	filan	itibaren	olan	senden	üstelik
ben	bunların	gene	kaç	olarak	seni	üzere
bence	bunu	gibi	kadar	oldukça	senin	var
beni	bunun	göre	kendi	olma	siz	ve
benim	burada	hala	kendine	olmak	sizden	veya
beri	bütün	halen	kendini	on	size	veyahut
beş	çoğu	hangi	ki	ona	sizi	ya
bile	çoğuna	hangisi	kim	ondan	sizin	yalnızca
bir	çoğunu	hani	kime	onlar	son	yani
birçoğu	çok	hatta	kimi	onlara	sonra	yapmak
birçok	çünkü	hem	kimin	onlardan	şayet	yedi
birçokları	da	henüz	kimisi	onları	şey	yerine
biri	daha	hep	kimse	onların	şeyden	yine
birisi	dahası	hepsi	madem	onu	şeye	yoksa
birkaç	de	hepsine	mı	onun	şeyi	zaten
birkaçı	değil	hepsini	mi	orada	şeyler	zira

Table A.1: Turkish stopwords list.

a	did	here	like	ourselves	them	we've
about	didnt	here's	long	out	themselves	we'd
above	do	hers	made	over	then	we'll
after	does	herself	make	own	there	wasn't
again	doesn't	high	many	put	there's	weren't
against	doing	him	may	said	these	won't
all	don't	himself	me	same	they	wouldn't
also	down	his	might	say	they'd	who's
am	during	how	more	says	they'll	what's
an	each	how's	most	second	they've	when's
and	even	however	must	see	they're	where's
another	ever	he's	mustn't	seen	this	why's
any	every	he'd	my	shall	those	while
are	few	he'll	myself	should	three	with
arent	first	i	never	she's	through	when
as	five	i'd	no	she	to	where
at	for	if	nor	she'd	too	why
back	four	i'll	not	she'll	two	very
be	from	i'm	now	shan't	under	well
because	further	in	new	shouldn't	until	way
been	get	into	of	since	up	you
before	go	is	off	so	whether	you'd
being	goes	isn't	old	some	we	you'll
below	had	it	on	still	what	your
between	hadn't	its	one	such	which	you're
both	has	it's	once	take	who	yours
but	hasn't	itself	only	than	whom	yourself
by	have	i've	or	that	was	yourselves
can	haven't	just	other	that's	were	you've
can't	having	least	ought	the	will	
cannot	he	less	our	their	would	
could	her	lets	ours	theirs	we're	

Table A.2: English stopwords list.

A.2 Turkish News RSS Feeds List

Bilkent News Portal gathers news in several different categories from 8 most popular Turkish news websites, which distribute frequently updated RSS feeds. The list of Turkish news RSS feeds is given below.

News Website	Category	URL of RSS Feed
CNN Türk	Bilişim	http://www.cnnturk.com/servisler/rss/bilim.teknoloji.rss
	Dünya	http://www.cnnturk.com/servisler/rss/dunya.rss
	Ekonomi	http://www.cnnturk.com/servisler/rss/ekonomi.rss
	Hava Durumu	http://www.cnnturk.com/servisler/rss/havadurumu.rss
	Kültür-Sanat	http://www.cnnturk.com/servisler/rss/kultur.sanat.rss
	Sağlık	http://www.cnnturk.com/servisler/rss/saglik.rss
	Spor	http://www.cnnturk.com/servisler/rss/spor.rss
	Türkiye	http://www.cnnturk.com/servisler/rss/turkiye.rss
	Yaşam	http://www.cnnturk.com/servisler/rss/yasam.rss
Hürriyet	Ana Sayfa	http://rss.hurriyet.com.tr/rss.aspx?sectionId=1
	Dünya	http://rss.hurriyet.com.tr/rss.aspx?sectionId=2249
	Ekonomi	http://rss.hurriyet.com.tr/rss.aspx?sectionId=4
	Kültür-Sanat	http://rss.hurriyet.com.tr/rss.aspx?sectionId=13
	Magazin	http://rss.hurriyet.com.tr/rss.aspx?sectionId=2035
	Sağlık	http://rss.hurriyet.com.tr/rss.aspx?sectionId=2208
	Spor	http://rss.hurriyet.com.tr/rss.aspx?sectionId=14
Milliyet	Dünya	http://www.milliyet.com.tr/D/rss/rss/Rss_2.xml
	Ekonomi	http://www.milliyet.com.tr/D/rss/rss/Rss_3.xml
	Sağlık	http://www.milliyet.com.tr/D/rss/rss/Rss_31.xml
	Siyaset	http://www.milliyet.com.tr/D/rss/rss/Rss_4.xml
	Spor	http://www.milliyet.com.tr/D/rss/rss/Rss_6.xml
	Teknoloji	http://www.milliyet.com.tr/D/rss/rss/Rss_36.xml
	Yaşam	http://www.milliyet.com.tr/D/rss/rss/Rss_5.xml
Sabah	Dünya	http://www.sabah.com.tr/rss/Dunya.xml
	Ekonomi	http://www.sabah.com.tr/rss/Ekonomi.xml
	Gündem	http://www.sabah.com.tr/rss/Gundem.xml
	Magazin	http://www.sabah.com.tr/rss/Magazin.xml
	Sağlık	http://www.sabah.com.tr/rss/Saglik.xml
	Spor	http://www.sabah.com.tr/rss/Spor.xml
	Teknoloji	http://www.sabah.com.tr/rss/Teknoloji.xml
	Yaşam	http://www.sabah.com.tr/rss/Yasam.xml
Star	Dünya	http://www.stargazete.com/dunya.xml
	Ekonomi	http://www.stargazete.com/ekonomi.xml
	Güncel	http://www.stargazete.com/guncel.xml
	Magazin	http://www.stargazete.com/rss/magazin.xml
	Politika	http://www.stargazete.com/politika.xml
	Sağlık	http://www.stargazete.com/rss/saglik.xml
	Sanat	http://www.stargazete.com/rss/sanat.xml
	Spor	http://www.stargazete.com/spor.xml
	Teknoloji	http://www.stargazete.com/rss/teknoloji.xml

Yeni Şafak	Bilişim	http://yenisafak.com.tr/rss/?xml=bilisim
	Gündem	http://yenisafak.com.tr/rss/?xml=gundem
	Dünya	http://yenisafak.com.tr/rss/?xml=dunya
	Ekonomi	http://yenisafak.com.tr/rss/?xml=ekonomi
	Kültür-Sanat	http://yenisafak.com.tr/rss/?xml=kultursanat
	Politika	http://yenisafak.com.tr/rss/?xml=politika
	Sağlk	http://yenisafak.com.tr/rss/?xml=saglik
	Spor	http://yenisafak.com.tr/rss/?xml=spor
Vatan	Dünya	http://rss.gazetevatan.com/rss/dunya.xml
	Ekonomi	http://rss.gazetevatan.com/rss/ekonomi.xml
	Gündem	http://rss.gazetevatan.com/rss/gundem.xml
	Magazin	http://rss.gazetevatan.com/rss/magazin.xml
	Siyaset	http://rss.gazetevatan.com/rss/siyaset.xml
	Spor	http://rss.gazetevatan.com/rss/spor.xml
	Teknoloji	http://rss.gazetevatan.com/rss/teknoloji.xml
	Yaşam	http://rss.gazetevatan.com/rss/yasam.xml
Zaman	Aile-Sağlık	http://www.zaman.com.tr/aile.rss
	Dış Haberler	http://www.zaman.com.tr/dishaberler.rss
	Ekonomi	http://www.zaman.com.tr/ekonomi.rss
	Gündem	http://www.zaman.com.tr/gundem.rss
	Kültür-Sanat	http://www.zaman.com.tr/kultursanat.rss
	Politika	http://www.zaman.com.tr/politika.rss
	Spor	http://www.zaman.com.tr/spor.rss

Table A.3: Turkish news RSS feeds list.

Appendix B

Calculation Examples

B.1 Similarity Calculation Examples

B.1.1 Vector Representations

Ali akşam eve geç geldi. | Babası, Ali'nin eve geç gelmesine kızdı. | Ali buna evde çok üzüldü.

Sentence #1 | Sentence #2 | Sentence #3

Location \ Term	Ali	akşam	ev	geç	gel	baba	gelme	kız	üzül
Document	3	1	3	2	1	1	1	1	1
Sentence #1	1	1	1	1	1	0	0	0	0
Sentence #2	1	0	1	1	0	1	1	1	0
Sentence #3	1	0	1	0	0	0	0	0	1

$$\text{Document Vector} = [\frac{3}{14}\ \frac{1}{14}\ \frac{3}{14}\ \frac{2}{14}\ \frac{1}{14}\ \frac{1}{14}\ \frac{1}{14}\ \frac{1}{14}\ \frac{1}{14}]$$

$$\text{Sentence \#1 Vector} = [\frac{1}{5}\ \frac{1}{5}\ \frac{1}{5}\ \frac{1}{5}\ \frac{1}{5}\ \frac{0}{5}\ \frac{0}{5}\ \frac{0}{5}\ \frac{0}{5}]$$

$$\text{Sentence \#2 Vector} = [\frac{1}{6}\ \frac{0}{6}\ \frac{1}{6}\ \frac{1}{6}\ \frac{0}{6}\ \frac{1}{6}\ \frac{1}{6}\ \frac{1}{6}\ \frac{0}{6}]$$

$$\text{Sentence \#3 Vector} = [\frac{1}{3}\ \frac{0}{3}\ \frac{1}{3}\ \frac{0}{3}\ \frac{0}{3}\ \frac{0}{3}\ \frac{0}{3}\ \frac{0}{3}\ \frac{1}{3}]$$

Figure B.1: Term frequency assignment and vector representation example.

B.1.2 Cosine Similarity Calculation Example

$$\text{Cos. Similarity (D, S}_1\text{)} = \frac{\left(\frac{3}{14}*\frac{1}{5}\right)+\left(\frac{1}{14}*\frac{1}{5}\right)+\left(\frac{3}{14}*\frac{1}{5}\right)+\left(\frac{2}{14}*\frac{1}{5}\right)+\left(\frac{1}{14}*\frac{1}{5}\right)+\left(\frac{1}{14}*\frac{0}{5}\right)+\left(\frac{1}{14}*\frac{0}{5}\right)+\left(\frac{1}{14}*\frac{0}{5}\right)+\left(\frac{1}{14}*\frac{0}{5}\right)}{\sqrt{\left(\frac{3}{14}\right)^2+\left(\frac{1}{14}\right)^2+\left(\frac{3}{14}\right)^2+\left(\frac{2}{14}\right)^2+\left(\frac{1}{14}\right)^2+\left(\frac{1}{14}\right)^2+\left(\frac{1}{14}\right)^2+\left(\frac{1}{14}\right)^2+\left(\frac{1}{14}\right)^2}*\sqrt{\left(\frac{1}{5}\right)^2+\left(\frac{1}{5}\right)^2+\left(\frac{1}{5}\right)^2+\left(\frac{1}{5}\right)^2+\left(\frac{1}{5}\right)^2+\left(\frac{0}{5}\right)^2+\left(\frac{0}{5}\right)^2+\left(\frac{0}{5}\right)^2+\left(\frac{0}{5}\right)^2}} = \mathbf{0.845}$$

$$\text{Cos. Similarity (D, S}_2\text{)} = \frac{\left(\frac{3}{14}*\frac{1}{6}\right)+\left(\frac{1}{14}*\frac{0}{6}\right)+\left(\frac{3}{14}*\frac{1}{6}\right)+\left(\frac{2}{14}*\frac{1}{6}\right)+\left(\frac{1}{14}*\frac{0}{6}\right)+\left(\frac{1}{14}*\frac{1}{6}\right)+\left(\frac{1}{14}*\frac{1}{6}\right)+\left(\frac{1}{14}*\frac{1}{6}\right)+\left(\frac{1}{14}*\frac{0}{6}\right)}{\sqrt{\left(\frac{3}{14}\right)^2+\left(\frac{1}{14}\right)^2+\left(\frac{3}{14}\right)^2+\left(\frac{2}{14}\right)^2+\left(\frac{1}{14}\right)^2+\left(\frac{1}{14}\right)^2+\left(\frac{1}{14}\right)^2+\left(\frac{1}{14}\right)^2+\left(\frac{1}{14}\right)^2}*\sqrt{\left(\frac{1}{6}\right)^2+\left(\frac{0}{6}\right)^2+\left(\frac{1}{6}\right)^2+\left(\frac{1}{6}\right)^2+\left(\frac{0}{6}\right)^2+\left(\frac{1}{6}\right)^2+\left(\frac{1}{6}\right)^2+\left(\frac{1}{6}\right)^2+\left(\frac{0}{6}\right)^2}} = \mathbf{0.849}$$

$$\text{Cos. Similarity (D, S}_3\text{)} = \frac{\left(\frac{3}{14}*\frac{1}{3}\right)+\left(\frac{1}{14}*\frac{0}{3}\right)+\left(\frac{3}{14}*\frac{1}{3}\right)+\left(\frac{2}{14}*\frac{0}{3}\right)+\left(\frac{1}{14}*\frac{0}{3}\right)+\left(\frac{1}{14}*\frac{0}{3}\right)+\left(\frac{1}{14}*\frac{0}{3}\right)+\left(\frac{1}{14}*\frac{0}{3}\right)+\left(\frac{1}{14}*\frac{1}{3}\right)}{\sqrt{\left(\frac{3}{14}\right)^2+\left(\frac{1}{14}\right)^2+\left(\frac{3}{14}\right)^2+\left(\frac{2}{14}\right)^2+\left(\frac{1}{14}\right)^2+\left(\frac{1}{14}\right)^2+\left(\frac{1}{14}\right)^2+\left(\frac{1}{14}\right)^2+\left(\frac{1}{14}\right)^2}*\sqrt{\left(\frac{1}{3}\right)^2+\left(\frac{0}{3}\right)^2+\left(\frac{1}{3}\right)^2+\left(\frac{0}{3}\right)^2+\left(\frac{0}{3}\right)^2+\left(\frac{0}{3}\right)^2+\left(\frac{0}{3}\right)^2+\left(\frac{0}{3}\right)^2+\left(\frac{1}{3}\right)^2}} = \mathbf{0.763}$$

Figure B.2: Calculation of the Cosine similarities of the example given in Figure B.1.

B.1.3 Dice Similarity Example

$$\text{Dice Coefficient (D, S}_1\text{)} = \frac{2*\left[\left(\frac{3}{14}*\frac{1}{5}\right)+\left(\frac{1}{14}*\frac{1}{5}\right)+\left(\frac{3}{14}*\frac{1}{5}\right)+\left(\frac{2}{14}*\frac{1}{5}\right)+\left(\frac{1}{14}*\frac{1}{5}\right)+\left(\frac{1}{14}*\frac{0}{5}\right)+\left(\frac{1}{14}*\frac{0}{5}\right)+\left(\frac{1}{14}*\frac{0}{5}\right)+\left(\frac{1}{14}*\frac{0}{5}\right)\right]}{\left(\frac{3}{14}^2+\frac{1}{14}^2+\frac{3}{14}^2+\frac{2}{14}^2+\frac{1}{14}^2+\frac{1}{14}^2+\frac{1}{14}^2+\frac{1}{14}^2+\frac{1}{14}^2\right)+\left(\frac{1}{5}^2+\frac{1}{5}^2+\frac{1}{5}^2+\frac{1}{5}^2+\frac{1}{5}^2+\frac{0}{5}^2+\frac{0}{5}^2+\frac{0}{5}^2+\frac{0}{5}^2\right)} = \mathbf{0.833}$$

$$\text{Dice Coefficient (D, S}_2\text{)} = \frac{2*\left[\left(\frac{3}{14}*\frac{1}{6}\right)+\left(\frac{1}{14}*\frac{0}{6}\right)+\left(\frac{3}{14}*\frac{1}{6}\right)+\left(\frac{2}{14}*\frac{1}{6}\right)+\left(\frac{1}{14}*\frac{0}{6}\right)+\left(\frac{1}{14}*\frac{1}{6}\right)+\left(\frac{1}{14}*\frac{1}{6}\right)+\left(\frac{1}{14}*\frac{1}{6}\right)+\left(\frac{1}{14}*\frac{0}{6}\right)\right]}{\left(\frac{3}{14}^2+\frac{1}{14}^2+\frac{3}{14}^2+\frac{2}{14}^2+\frac{1}{14}^2+\frac{1}{14}^2+\frac{1}{14}^2+\frac{1}{14}^2+\frac{1}{14}^2\right)+\left(\frac{1}{6}^2+\frac{0}{6}^2+\frac{1}{6}^2+\frac{1}{6}^2+\frac{0}{6}^2+\frac{1}{6}^2+\frac{1}{6}^2+\frac{1}{6}^2+\frac{0}{6}^2\right)} = \mathbf{0.846}$$

$$\text{Dice Coefficient (D, S}_3\text{)} = \frac{2*\left[\left(\frac{3}{14}*\frac{1}{3}\right)+\left(\frac{1}{14}*\frac{0}{3}\right)+\left(\frac{3}{14}*\frac{1}{3}\right)+\left(\frac{2}{14}*\frac{0}{3}\right)+\left(\frac{1}{14}*\frac{0}{3}\right)+\left(\frac{1}{14}*\frac{0}{3}\right)+\left(\frac{1}{14}*\frac{0}{3}\right)+\left(\frac{1}{14}*\frac{0}{3}\right)+\left(\frac{1}{14}*\frac{1}{3}\right)\right]}{\left(\frac{3}{14}^2+\frac{1}{14}^2+\frac{3}{14}^2+\frac{2}{14}^2+\frac{1}{14}^2+\frac{1}{14}^2+\frac{1}{14}^2+\frac{1}{14}^2+\frac{1}{14}^2\right)+\left(\frac{1}{3}^2+\frac{0}{3}^2+\frac{1}{3}^2+\frac{0}{3}^2+\frac{0}{3}^2+\frac{0}{3}^2+\frac{0}{3}^2+\frac{0}{3}^2+\frac{1}{3}^2\right)} = \mathbf{0.7}$$

Figure B.3: Calculation of the Dice similarities of the example given in Figure B.1.

B.1.4 Jaccard Similarity Example

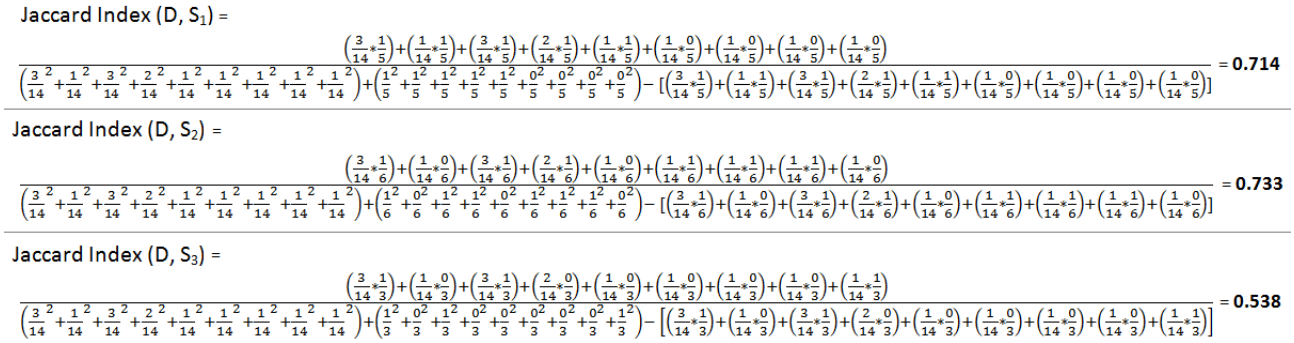


Figure B.4: Calculation of the Jaccard similarities of the example given in Figure B.1.

B.1.5 Overlap Similarity Example

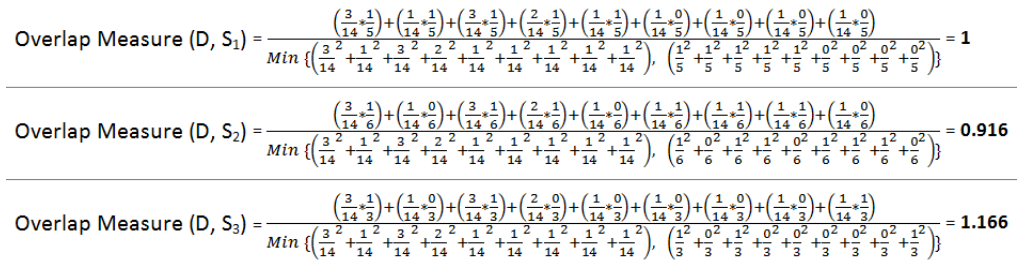


Figure B.5: Calculation of the Overlap similarities of the example given in Figure B.1.

B.2 Means Comparison Calculation Examples

B.2.1 ANOVA Calculation Example

Groups	Test Scores							
1	7	4	6	8	6	6	2	9
2	5	5	3	4	4	7	2	2
3	2	4	7	1	2	1	5	5

x_1	x_1^2	x_2	x_2^2	x_3	x_3^2
7	49	5	25	2	4
4	16	5	25	4	16
6	36	3	9	7	49
8	64	4	16	1	1
6	36	4	16	2	4
6	36	7	49	1	1
2	4	2	4	5	25
9	81	2	4	5	25
$\Sigma x_1 = 48$	$\Sigma x_1^2 = 322$	$\Sigma x_2 = 32$	$\Sigma x_2^2 = 148$	$\Sigma x_3 = 27$	$\Sigma x_3^2 = 125$
$(\Sigma x_1)^2 = 2304$		$(\Sigma x_2)^2 = 1024$		$(\Sigma x_3)^2 = 729$	
$M_1 = 6$		$M_2 = 4$		$M_3 = 3.375$	

$$SS_{total} = (322 + 148 + 125) - \frac{(48 + 32 + 27)^2}{24} = 595 - 477.04 = 117.96$$

$$SS_{among} = \left(\frac{2304}{8} + \frac{1024}{8} + \frac{729}{8}\right) - 477.04 = 507.13 - 477.04 = 30.08$$

$$SS_{within} = 117.96 - 30.08 = 87.88$$

$$df_{among} = 3 - 1 = 2 \qquad df_{within} = 24 - 3 = 21$$

$$MS_{among} = \frac{30.08}{2} = 15.04 \qquad MS_{within} = \frac{87.88}{21} = 4.18$$

$$F_{calculated} = \frac{15.04}{4.18} = 3.59 \qquad F_{table} = 3.4668$$

$F_{calculated} > F_{table}$ => F score is **statistically significant.**

Figure B.6: ANOVA calculation example.

B.2.2 Scheffé's Test Calculation Example

MS_{within} =4.18, M_1 =6, M_2 =4, M_3 =3, df_{within} = 21, df_{among} = 2

n_1 = 8, n_2 = 8, k = 3, $F_{table(2,21)}$ = 3.4668

$$\boldsymbol{F_{critical}} = (3-1) \times 3.4668 = 6.9336$$

$$F_{1,2} = \frac{(6-4)^2}{4.18 \times (\frac{1}{8}+\frac{1}{8})} = 3.828 \Rightarrow \boldsymbol{F_{1,2}} < \boldsymbol{F_{critical}}$$

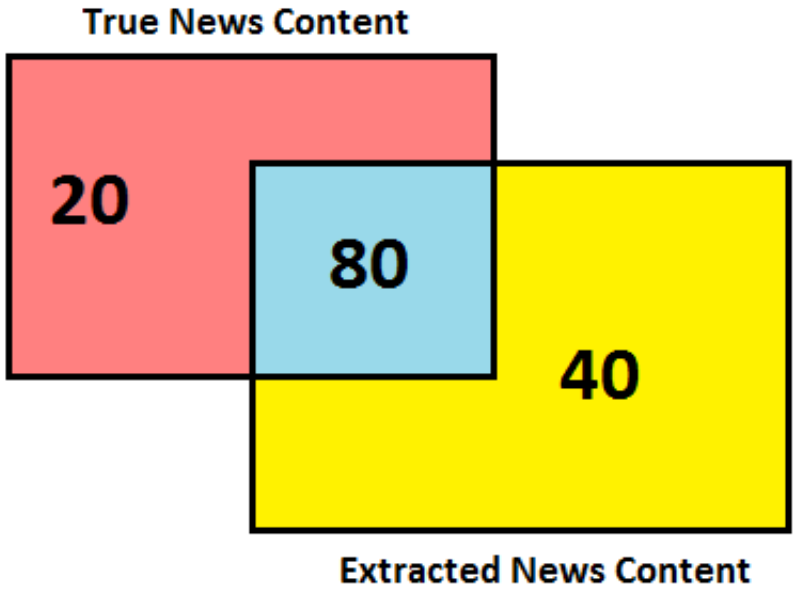
$$F_{1,3} = \frac{(6-3)^2}{4.18 \times (\frac{1}{8}+\frac{1}{8})} = 8.612 \Rightarrow \boldsymbol{F_{1,3}} > \boldsymbol{F_{critical}}$$

$$F_{2,3} = \frac{(4-3)^2}{4.18 \times (\frac{1}{8}+\frac{1}{8})} = 0.957 \Rightarrow \boldsymbol{F_{2,3}} < \boldsymbol{F_{critical}}$$

Hence, only means **1** and **3** are significantly different from each other.

Figure B.7: Scheffé's test calculation example.

B.3 Set-based Measures Calculation Example



|True News Content| = 100
|Extracted News Content| = 120
|True Positive (Relevatn & Extracted)| = 80
|False Negative (Relevant & Not Extracted)| = 20
|False Positive (Irrelevant & Extracted)| = 40

$$Precision = \frac{80}{120} = 0.667$$

$$Recall = \frac{80}{100} = 0.8$$

$$F - measure = 2 \times \frac{(0.667 \times 0.8)}{(0.667 + 0.8)} = 0.727$$

Figure B.8: Set-based measures calculation example.

Appendix C

Additional Experimental Results Using Cosine, Jaccard, and Overlap Similarity Measures

C.1 Additional NBD Results

News block detection (NBD) accuracy training and testing results of N-EXT with TR-Block and ENG-Block datasets using other similarity measures (Cosine, Jaccard, and Overlap) are given below.

β / k	β Values										
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
1	0.901	0.907	0.911	0.916	0.920	0.925	**0.930**	0.924	0.919	0.913	0.905
2	0.899	0.905	0.909	0.916	0.923	0.927	**0.932**	0.929	0.925	0.918	0.910
3	0.902	0.904	0.907	0.911	0.920	0.923	0.926	**0.929**	0.927	0.921	0.912
4	0.895	0.903	0.906	0.909	0.911	0.915	**0.921**	0.919	0.917	0.912	0.906
5	0.908	0.910	0.914	0.918	0.920	**0.925**	**0.925**	0.922	0.919	0.914	0.911
6	0.902	0.908	0.910	0.915	0.921	0.928	**0.929**	0.927	0.923	0.920	0.913
7	0.906	0.908	0.911	0.913	0.917	0.921	**0.927**	0.924	0.920	0.915	0.908
8	0.901	0.905	0.909	0.912	0.914	0.921	0.925	**0.929**	0.923	0.920	0.914
9	0.900	0.904	0.907	0.912	0.916	0.921	**0.922**	0.921	0.916	0.915	0.910
10	0.904	0.910	0.915	0.919	0.923	0.927	**0.932**	0.930	0.927	0.926	0.918
Avg.	0.902	0.906	0.910	0.914	0.918	0.923	**0.927**	0.925	0.922	0.918	0.911

a) NBD accuracy training results.

β / k	β Values										
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
1	0.87	0.88	0.89	0.89	0.91	0.92	**0.93**	0.92	0.91	0.91	0.89
2	0.85	0.87	0.88	0.89	0.91	0.91	**0.93**	0.91	0.90	0.90	0.88
3	0.86	0.87	0.89	0.90	0.90	0.91	0.92	**0.93**	0.91	0.91	0.90
4	0.85	0.87	0.88	0.90	0.91	0.91	**0.92**	0.90	0.89	0.88	0.87
5	0.85	0.87	0.88	0.88	0.90	**0.91**	**0.91**	0.90	0.88	0.87	0.87
6	0.87	0.89	0.91	0.91	0.91	0.92	**0.93**	0.92	0.90	0.89	0.88
7	0.88	0.90	0.90	0.90	0.91	0.91	**0.93**	0.91	0.91	0.90	0.89
8	0.86	0.87	0.89	0.89	0.91	**0.92**	**0.92**	**0.92**	0.90	0.90	0.88
9	0.87	0.88	0.90	0.90	0.91	**0.92**	**0.92**	0.91	0.90	0.89	0.87
10	0.89	0.90	0.91	0.91	0.91	0.93	**0.94**	0.93	0.91	0.90	0.88
Avg.	0.865	0.880	0.893	0.897	0.908	0.916	**0.925**	0.915	0.903	0.895	0.881

b) NBD accuracy testing results.

Table C.1: News block detection (NBD) accuracy training and testing results of N-EXT with TR-Block dataset (without hyperlink texts) using Cosine similarity measure and 10-fold cross-validation.

β / k	β Values										
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
1	0.919	0.926	0.933	0.941	0.945	0.953	**0.961**	0.952	0.944	0.933	0.924
2	0.923	0.931	0.936	0.939	0.947	0.954	**0.958**	0.947	0.939	0.930	0.921
3	0.922	0.929	0.935	0.938	0.943	0.948	0.950	**0.952**	0.943	0.932	0.925
4	0.918	0.923	0.935	0.941	0.945	0.949	**0.956**	0.945	0.938	0.929	0.919
5	0.921	0.925	0.934	0.938	0.947	**0.954**	**0.954**	0.948	0.944	0.935	0.924
6	0.922	0.933	0.937	0.944	0.950	0.955	**0.962**	0.953	0.944	0.936	0.926
7	0.917	0.925	0.933	0.942	0.950	0.954	**0.959**	0.946	0.938	0.929	0.920
8	0.923	0.930	0.934	0.939	0.946	0.952	**0.955**	0.954	0.946	0.935	0.924
9	0.918	0.924	0.931	0.938	0.947	0.952	0.952	**0.954**	0.945	0.938	0.929
10	0.920	0.925	0.933	0.942	0.948	0.956	**0.963**	0.954	0.944	0.931	0.922
Avg.	0.920	0.927	0.934	0.940	0.947	0.953	**0.957**	0.951	0.943	0.934	0.923

a) NBD accuracy training results.

β / k	β Values										
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
1	0.90	0.91	0.92	0.92	0.94	0.94	**0.96**	0.94	0.94	0.92	0.91
2	0.91	0.92	0.92	0.94	0.94	0.95	**0.96**	0.95	0.94	0.92	0.90
3	0.90	0.91	0.92	0.92	0.92	0.94	0.95	**0.96**	0.94	0.93	0.91
4	0.90	0.91	0.93	0.93	0.93	0.94	**0.95**	0.93	0.93	0.92	0.90
5	0.89	0.90	0.92	0.92	0.93	0.93	**0.94**	0.93	0.93	0.91	0.90
6	0.90	0.91	0.93	0.93	0.95	0.95	**0.96**	0.95	0.94	0.92	0.92
7	0.90	0.91	0.92	0.92	0.94	0.94	**0.96**	**0.96**	0.94	0.92	0.90
8	0.89	0.90	0.91	0.93	0.93	0.94	**0.95**	0.94	0.92	0.92	0.90
9	0.90	0.90	0.92	0.92	0.93	**0.94**	**0.94**	0.92	0.92	0.92	0.91
10	0.92	0.92	0.93	0.93	0.94	0.94	**0.96**	0.95	0.93	0.92	0.90
Avg.	0.901	0.909	0.922	0.926	0.935	0.941	**0.953**	0.943	0.933	0.920	0.905

b) NBD accuracy testing results.

Table C.2: News block detection (NBD) accuracy training and testing results of N-EXT with TR-Block dataset (without hyperlink texts) using Jaccard similarity measure and 10-fold cross-validation.

β / k	β Values										
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
1	0.922	0.928	0.936	0.940	0.909	0.915	**0.919**	0.951	0.942	0.929	0.918
2	0.919	0.926	0.937	0.941	0.949	0.958	**0.967**	0.950	0.949	0.924	0.912
3	0.925	0.929	0.939	0.941	0.952	0.954	0.955	**0.957**	0.948	0.932	0.921
4	0.923	0.926	0.937	0.943	0.950	0.954	**0.961**	0.952	0.947	0.930	0.922
5	0.921	0.925	0.934	0.938	0.947	**0.955**	0.954	0.952	0.946	0.931	0.924
6	0.923	0.932	0.939	0.946	0.954	0.958	**0.969**	0.959	0.948	0.938	0.928
7	0.924	0.927	0.935	0.943	0.949	0.955	**0.963**	0.951	0.942	0.929	0.918
8	0.924	0.931	0.938	0.942	0.944	0.951	0.953	**0.956**	0.947	0.933	0.925
9	0.923	0.926	0.938	0.945	0.949	0.952	0.954	**0.955**	0.946	0.939	0.931
10	0.925	0.929	0.936	0.945	0.947	0.956	**0.968**	0.953	0.944	0.927	0.920
Avg.	0.923	0.928	0.937	0.942	0.949	0.955	**0.961**	0.954	0.946	0.931	0.922

a) NBD accuracy training results.

β / k	β Values										
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
1	0.85	0.87	0.88	0.89	0.90	0.90	**0.91**	0.89	0.88	0.86	0.86
2	0.85	0.86	0.88	0.88	0.88	0.90	**0.91**	0.89	0.89	0.87	0.85
3	0.85	0.87	0.87	0.88	0.89	0.89	**0.90**	**0.90**	0.88	0.86	0.85
4	0.84	0.86	0.86	0.88	0.90	0.90	**0.91**	0.90	0.89	0.87	0.87
5	0.85	0.87	0.87	0.88	0.88	**0.90**	**0.90**	0.88	0.87	0.86	0.84
6	0.84	0.86	0.87	0.87	0.89	0.90	**0.92**	0.91	0.90	0.88	0.87
7	0.85	0.86	0.86	0.88	0.88	0.89	**0.91**	0.89	0.89	0.87	0.86
8	0.86	0.86	0.88	0.89	0.91	0.91	**0.92**	0.90	0.90	0.88	0.87
9	0.85	0.85	0.87	0.88	0.90	0.90	**0.91**	0.90	0.89	0.88	0.88
10	0.84	0.85	0.87	0.89	0.91	0.91	**0.93**	0.92	0.90	0.90	0.89
Avg.	0.848	0.861	0.871	0.882	0.894	0.900	**0.911**	0.898	0.889	0.873	0.864

b) NBD accuracy testing results.

Table C.3: News block detection (NBD) accuracy training and testing results of N-EXT with TR-Block dataset (without hyperlink texts) using Overlap similarity measure and 10-fold cross-validation.

β / ht	β Values										
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
With Hyperlink Texts	0.75	0.76	0.76	0.77	0.77	0.79	**0.80**	0.79	0.77	0.77	0.76
Without Hyperlink Texts	0.83	0.84	0.85	0.85	0.87	0.88	**0.89**	0.87	0.87	0.86	0.84

Table C.4: News block detection (NBD) accuracy results of N-EXT with ENG-Block dataset using Cosine similarity measure.

β / ht	β Values										
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
With Hyperlink Texts	0.77	0.78	0.79	0.79	0.80	0.80	**0.82**	0.80	0.80	0.78	0.78
Without Hyperlink Texts	0.86	0.86	0.88	0.88	0.89	0.89	**0.91**	0.90	0.88	0.88	0.87

Table C.5: News block detection (NBD) accuracy results of N-EXT with ENG-Block dataset using Jaccard similarity measure.

β / ht	β Values										
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
With Hyperlink Texts	0.72	0.74	0.75	0.75	0.75	**0.77**	**0.77**	0.76	0.76	0.75	0.73
Without Hyperlink Texts	0.81	0.83	0.83	0.83	0.84	0.85	**0.86**	0.85	0.84	0.84	0.83

Table C.6: News block detection (NBD) accuracy results of N-EXT with ENG-Block dataset using Overlap similarity measure.

Datasets		Approaches		
		Only Similarity ($\boldsymbol{\beta} = 0$)	**Size & Similarity** ($\boldsymbol{\beta} = 0.6$)	**Only Size** ($\boldsymbol{\beta} = 0$)
Turkish	with hyperlink texts	0.802	0.831	0.805
	w/o hyperlink texts	0.865	**0.925**	0.881
English	with hyperlink texts	0.750	0.800	0.760
	w/o hyperlink texts	0.830	**0.890**	0.840

Table C.7: Summary of the news block detection (NBD) accuracy results of N-EXT with TR-Block and ENG-Block datasets using Cosine similarity measure.

Datasets		Approaches		
		Only Similarity ($\boldsymbol{\beta} = 0$)	**Size & Similarity** ($\boldsymbol{\beta} = 0.6$)	**Only Size** ($\boldsymbol{\beta} = 0$)
Turkish	with hyperlink texts	0.815	0.849	0.809
	w/o hyperlink texts	0.901	**0.953**	0.905
English	with hyperlink texts	0.770	0.820	0.780
	w/o hyperlink texts	0.860	**0.910**	0.870

Table C.8: Summary of the news block detection (NBD) accuracy results of N-EXT with TR-Block and ENG-Block datasets using Jaccard similarity measure.

Datasets		Approaches		
		Only Similarity ($\boldsymbol{\beta} = 0$)	**Size & Similarity** ($\boldsymbol{\beta} = 0.6$)	**Only Size** ($\boldsymbol{\beta} = 0$)
Turkish	with hyperlink texts	0.792	0.824	0.799
	w/o hyperlink texts	0.848	**0.911**	0.864
English	with hyperlink texts	0.720	0.770	0.730
	w/o hyperlink texts	0.810	**0.860**	0.830

Table C.9: Summary of the news block detection (NBD) accuracy results of N-EXT with TR-Block and ENG-Block datasets using Overlap similarity measure.

C.2 Additional NCE Results

News Websites	Similarity Measures				Average
	Cosine	Dice	Jaccard	Overlap	
CNN Türk	0.890	0.915	0.911	0.884	0.900
Milliyet	0.886	0.906	0.902	0.888	0.896
Sabah	0.887	0.912	0.908	0.891	0.899
Samanyolu	0.901	0.915	0.910	0.895	0.905
Star	0.906	0.921	0.919	0.903	**0.912**
Yeni Şafak	0.902	0.914	0.911	0.901	0.907
Zaman	0.904	0.914	0.909	0.899	0.906
Average	0.897	**0.914**	0.910	0.894	0.904

Table C.10: Average F-measure values for different news websites without using stemming.