News Item Extraction for Text Mining in Web Newspapers

Kjetil Nørvåg* and Randi Øyri
Department of Computer and Information Science
Norwegian University of Science and Technology
7491 Trondheim, Norway

Abstract

Web newspapers provide a valuable resource for information. In order to benefit more from the available information, text mining techniques can be applied. However, because each newspaper page often covers a lot of unrelated topics, page-based data mining will not always give useful results. In order to improve on complete-page mining, we present an approach based on extracting the individual news items from the web pages and mining these separately. Automatic news item extraction is a difficult problem, and in this paper we also provide strategies solving that task. We study the quality of the news item extraction, and also provide results from clustering the extracted news items.

1 Introduction

During the last decade, most major newspapers and magazines have developed web sites providing news or other material. In addition, web-only newspapers have also appeared. The quality as well as the amount of what is presented on all these web sites have considerably improved, thus providing a valuable resource for information.

Information on the newspaper web sites have for some time been easily searchable through site-specific search tools, as well as popular search engines like Google, Yahoo, etc. These essentially provide keyword-based search, although some of the search engines also provide clustering-based tools for finding related pages. However, these are still relatively primitive, and in the context of web newspapers they are not always very useful: because each newspaper page often covers a lot of unrelated topics, page-based clustering will not give useful results.

In order to improve on complete-page mining, our approach is based on extracting the individual news items from the web pages and mining these separately. This should considerably increase the quality of the results, because 1) these news items are short, 2) contain relevant and descriptive key words, and 3) are made by humans which in general gives a higher quality of the classification compared to what would be the result of using automatic techniques to do the task. Our approach also has another very important advantage: The news items on the entry pages of a newspaper essentially provides us with a compressed version of the full stories, so that by only performing the mining on the main pages of the web sites the amount of data to mine is reduced. The removal of non-relevant information also makes news item extraction useful as a way of data cleaning.

Although news item extraction is a relatively easy task for a human who can do it just by visual inspection, it is a hard problem for a computer, and previous approaches to the problem have not been robust enough. In this paper, we present a pattern-based strategy that we show provides a satisfactory quality as well as being robust in the sense that it performs well for different web newspapers.

The web pages are dynamic and changes continuously as news is published. In order to capture the information that is published on these pages, we retrieve the main pages of the web newspapers at regular intervals, and store all the retrieved versions. The stored versions can also be utilized for other purposes, and this work is done in the context of the V2 temporal document database system [14].

The mining process using our approach is based on a repository of web newspaper pages and extracting items from these pages (in this case, a news item is identified by the combination of the URL of web page, the timestamp of the web page version and an identifier for each news item on the page). The news items are then used in the data mining process. We will in this paper present an example of clustering news items and the results achieved. Other data mining techniques, for example mining association rules, can also be applied on the news items. A news item usually provides a link to the full story, this can be stored together with the news item, so that it is possible to access the relevant full stories after the data mining process.

The contributions of this paper are 1) news item based

*Email of contact author: Kjetil.Norvag@idi.ntnu.no

data mining instead of page-based mining, and 2) strategies for news item extraction and a study of the quality of these strategies.

The organization of the rest of this paper is as follows. In Section 2 we give an overview of related work. In Section 3 we describe how to extract news items from web pages. In Section 4 we study the quality of the news item extraction when our pattern-based strategy is used. In Section 5 we describe one of the data mining experiments we have performed on extracted news items. Finally, in Section 6, we conclude the paper.

2 Related work

Web page information extraction. Yi et al. describe in [18] how to remove irrelevant information in web pages in order to increase the quality of subsequent data mining. Their goal is to remove advertisements, navigation fields, copyright information, etc. This is achieved by detecting common elements in different pages belonging to the same site. Compared to our work, they do not attempt to detect individual news items on the pages. Other approaches with similar goals are proposed by Bar-Yossef and Rajagopalan in [1] and by Ramaswamy et al. in [15].

In [10] and [13] Kao et al. and Lin and Ho present methods to extract informative information from web page tables (<TABLE> in HTML).

An approach to detect content structure on web pages based on visual representation was presented by Cai et al. [3]. The approach is based on heuristics to process DOM trees. The author states that the experiments show satisfactory results. However, the approach does not handle web pages where images (for example thin lines) are used as separators. This is an important problem, because images are commonly used for this purpose (see, e.g., `http://www.vg.no`).

Embley et al. [7] present heuristics for extracting "records" from web pages. However, their approach is domain specific and requires an ontology for each domain.

News engines. Most of the well-known search engines like Google and Yahoo also extract information from web pages and categorize them according to topic. We assume that these companies have developed some extraction algorithms, but how these work are not publically available.

Wrappers. According to Laender et al. [11], the traditional method to extract information from web pages is to develop wrappers. The wrapper takes as input a web page containing an amount of information, and creates a mapping from the page to another format, for example news items. This mapping can be used in subsequent processing of similar pages, for example new versions of the web page. An example of a wrapper generator is W4F [17]. One problem with wrappers is that they require user interaction, and are less useful for large amounts of web pages.

A system similar to a wrapper, but that does not require training data, is presented by Chang et al. [4].

Data/text mining. Introduction to data mining in general and common data clustering algorithms can be found in many good text books, for example [8].

Various data mining techniques have been applied to text documents. We will in this paper restrict the discussion to clustering. Clustering can be used to find related documents. This can be used in a user-guided search process, but is also useful in other contexts. For example, Broder et al. [2] applied clustering to web documents, in order to find pairs or clusters of web pages that were (almost) identical. The result could be used for copy detection as well as finding lost pages. Clustering can also be used to discover topic hierarchies in large document collections [12].

An important issue in the context of clustering of text document, is dimensionality reduction, see, e.g., [6]. In Dhillon et al. [5] a complete framework for text clustering is given, providing techniques (and software) for the whole document clustering process.

There also exist commercial products providing support for text mining. One example is *IBM Intelligent Miner for Text* [9].

3 Extracting news items

For many types of web pages it is possible to use individual pages as the basis for data mining. However, in the case of the main pages of web newspapers this will often not give satisfactory results. For example, because each page covers a lot of unrelated topics, clustering will not give useful results (although this kind of clustering can be used to identify other similarities, for example pages on the same web site but using different URLs).[1] An alternative to clustering the pages containing the full stories is to use keywords as classification and perform clustering based on these keywords. However, classification can be a costly task in itself, in addition to problems with achieving high accuracy if performed without human cooperation.

Our solution is to use the short description of the stories that are available on the front pages as basis for the mining process. The description is short, and contains relevant and descriptive key words. An important advantage is that these are made by humans, in general giving a high accuracy as

[1] It should be noted that because the pages containing the full story also contain advertisements and links to other parts of the newspaper, news item extraction should be employed on these as well if they are to be used in data mining.

classifying words compared to what would be the result of using automatic techniques to perform the task.

The main page of a web newspaper normally consists of a number of *news items*. A news item on the main page is normally a short version of a larger news story, and the news story itself can normally be accessed by following a provided link. In order to improve the quality of the data mining process, we need to extract the individual news items on the web pages. For the rest of this paper we denote a news item to be the combination of the text and the associated URL that points to the full story (when known).

A web page can in general be considered as a collection of fragments. In the context or newspaper web pages, only some of the fragments are news items that are of interest to us. Other fragments that should not be considered as news items should be removed. Examples are:

- General information about the newspaper, for example information about the editors.
- Advertisements.
- Links/anchor-text to regular columns like weather, sports, etc.

Thus, our goal is to extract the news items as well as the links to the full stories. The quality of extracted news items is important for the subsequent data mining process, and this means that the extracted news items should be as complete as possible (for example, avoid that a news item as seen by a reader being identified as two items by the extraction tool), and at the same time avoid fragments in the three categories above being characterized as news items and contaminating the results.

It should be mentioned that developers of the web pages often by intension try to make the task of automatic extraction harder, in order to avoid advertisement-blocking tools to succeed. We can also see that different countries have developed different traditions on how a newspaper web page looks like and is constructed (for example, in some countries like Taiwan most newspaper web pages seem to be based on tables).

We will now present our approach to news item extraction and outline the full extraction process.

3.1 Pattern-based news item extraction

A person views the visual representation of a web page through a browser. For the person, it is a relatively easy task to determine what are the news items on the page. However, it is more difficult for a computer to extract the news items. Although it would also be possible for a computer to render the web page and try to use image processing techniques to extract items just like humans do, this would be a resource-demanding process, and not necessarily resulting in high accuracy. Our approach to solve the task, is to use simple strategies based on analyzing the HTML code of the web pages.

What the human sees as a well-defined item, can be very difficult to discover in the HTML code. Our strategies are based on detecting news item patterns that frequently occur in newspaper web pages: URL-text-URL item, line item, anchor-text item, bold header item, and text-based item. These patterns have been discovered by manually inspecting a large number of newspapers on the web and they will now be described in more detail. As will be shown in Section 4.2, a strategy based on detecting these patterns is sufficient to achieve sufficient quality.

3.1.1 URL-text-URL item

A common and easy-to-detect pattern is a heading that includes a link (to the full story), followed by some text, and then followed by the same link to the full story. In many cases, one of the links are associated with an image, as illustrated in Figure 1.

An example of a newspaper that frequently use this pattern is Dagbladet.[2] Another newspaper that uses this pattern for parts of the page is Washington Post,[3] where one of the links is associated with an image.

3.1.2 Line item

Some newspapers have each news item on a separate line in the code. An example of a site using this pattern is `http://www.dagbladet.no`.

In practice, the line item strategy is too specific and not sufficiently robust. It will be difficult to identify which sites actually follow this pattern, and manual interaction s likely needed. Therefore, we will not consider this pattern further.

3.1.3 Anchor-text item

On a web page, anchor text is the text that is visible in a link to a web page, i.e., the text between <A HREF=...> and </A>. In some newspapers on the web, the whole news items are contained in the anchor text. These anchor texts can be relatively long (more than 30 words) compared to typical links which only contain one or a few words.

A complicating factor when considering anchor-text items, is that many anchor texts should not be considered as news items. One example is links to subpages, for example *Sports* and *Weather*. Another example is links that are simply advertisements, although advertisements more often use images than anchor text. We employ two strategies to filter out subpage links (and other irrelevant anchor texts as well) that should not be considered as news items:

[2] `http://www.dagbladet.no`
[3] `http://www.washingtonpost.com/`


```
 <A HREF =
"http://go.vg.no/cgi-bin/front.cgi/0/http://www.vg.no/pub/vgart.hbs?artid=230147"> <IMG
SRC="http://blade.vg.no/grafikk/frontbilder/1356.jpg" BORDER="0" WIDTH="81" HEIGHT="100" ALT="Foto:  NRK1"
HSPACE="3" vspace="2"> </A> </span> <span CLASS="f-title-normal"> Tidenes heteste country-ballader
<BR> </span> </span> <A HREF =
"http://go.vg.no/cgi-bin/front.cgi/0/ http://www.vg.no/pub/vgart.hbs?artid=230147" CLASS="f-link"
TARGET="">Se listen her!  </A>
```

Figure 1. URL-text-URL item with image as first link. To the left is shown how the text and the image look on the web page, to the right the HTML code of the item is shown.

- Only consider texts larger than a certain number of words as news items. A suitable threshold value can be in the order of 5 words. Although a few news items with less text than 5 words can be left out by doing this, such small texts would in any case often be of little value in the data mining process.
- Our application context is a web warehouse that keeps all previously retrieved versions of the web pages. The time between each web page version is relatively short (typically one hour), so a news item will often be part of more than one version. In order to improve the quality of the data mining process, duplicates are removed so that only one copy of a news item is kept. A side effect of this processing is that also recurring anchor texts (for example the ones that points to the sport section) are removed.

Examples of web newspapers almost exclusively using anchor-text news items are the French newspaper Le Monde [4] and the Danish B.T.[5] In addition, the links on many other sites can be considered anchor-text items, but in that case only a few words are used in each item. One example is The New York Times [6] that have a few large news items at the top, followed by a large number of small anchor-text items further down.

In the case of anchor-text items there is a risk that there is additional text that is not part of the anchor text and that is omitted during the extraction process.

3.1.4 Bold header item

Some newspapers do not follow the patterns above, but start a news item with a header in bold font as showed in Figure 2. In this example, the whole body of the news item is just “Les hele saken” (Norwegian for “Read the whole story”), and it is the header that actually contains the meaningful words. In general, a news item in a normal font follows the header, and a link is given at the start or end of the text.

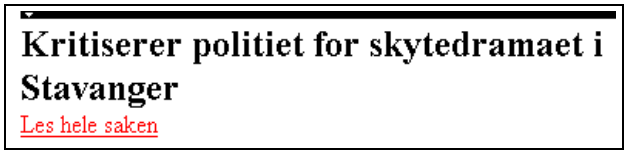


Figure 2. Example of emphasized news item from vg.no.

A complicating factor when analyzing a web page is that text on a web page can be set in bold in a number of ways, for example by the use of: <B> (bold), <H1> to <H6> (header styles), <BIG> (large font),<FONT SIZE> (different font sizes), <STRONG> (strong emphasis), and <SPAN CLASS> (size defined in separate stylesheets). Very often it is difficult to determine if text is in bold without processing a separate CSS-file. This can considerably increase the cost of detecting news items using this pattern.

3.1.5 Text-based item

In general, if there is a relatively long text with 1) no tags in between, or 2) only format tags like <P>, or 3) certain other tags like links for terms, it is very likely that the text is a news item. In many newspapers on the web, it is very common with one particular news item/story presented at the top. One such example is New York Post, and an example is given in Figure 3. By also allowing simple links inside text like the news item from The Chronicle of Higher Education[7] illustrated in Figure 4, we avoid news items be-

[4] http://www.lemonde.fr/
[5] http://www.bt.dk/
[6] http://www.times.com/
[7] http://chronicle.com/

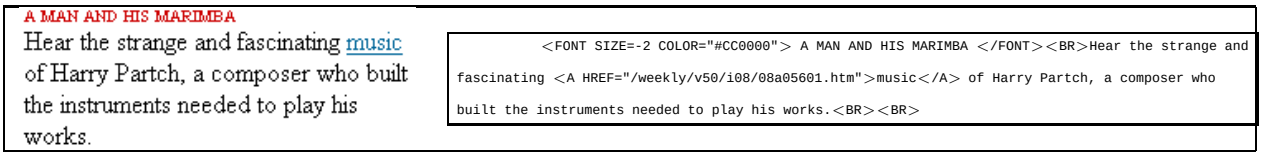


Figure 4. Typical news item from `chronicle.com`.

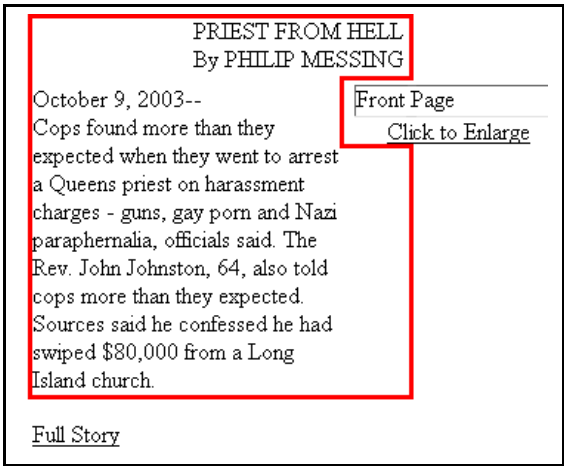


Figure 3. Text-based item from `nypost.com`. The text inside the border does not contain any links and is considered to be a news item. The link to the full story is further down.

ing broken up in pieces.

A possible problem with detecting a text-based item is that the link associated with the news item can be difficult to determine: it can be difficult to know whether it is the link before or after the text that links to the full story.

3.2 Extraction process

During the extraction process, 1) a web page is taken as input, 2) the web page is parsed and an attempt is made to detect the patterns as previously described, and 3) a number of news items (text/URL tuples) are produced as output. Some simple data cleaning operations are also performed on the news items in order to increase their quality with respect to data mining.

News items can fit into more than one pattern. During automatic extraction of news items, the initial strategy of our extraction tool is to detect items following the URL-text-URL pattern. This pattern has two identical links that function as news item delimiters, making the detection accurate. The tool also detects items following the anchor text and text-based patterns. Figure 5 illustrates some of the results from extracting news items from the Norwegian newspaper Dagens Næringsliv.[8]

4 Study of web pages and news item extraction quality

In this section we will study web page structure with respect to the patterns previously described, and study the quality of our data extraction approach.

In order to verify the quality and robustness of the pattern-based news item extraction approach as described above, we have compared the results of our news item extraction tool with manual inspection of a number of web pages. In our data mining experiments we have used a very large number of web pages, and because the manual inspection is a time consuming process we have performed the comparison of a subset of the web pages. We know that different countries and societies tend to have different structure and layout on web pages, so we have been careful to choose a subset with different characteristics: both local, regional and national newspapers, both main pages and subpages (for example, sports and entertainment), and different languages. Due to space constraints we only show results from a few of the inspected pages in this paper. The results from the comparison of the outpur from our tool with other pages was comparable to those presented here.

4.1 Web page structure

In order to study to what extent the various patterns occur in newspapers on the web, we employed the extraction tool on a number of web pages. Figure 6(a) illustrates the distribution of extracted news items between different patterns. As expected from what we have previously seen, the occurrence of the different patterns vary very much between different newspapers. While some have a very skewed distribution, like `chronicle.com` and `sognavis.no`, most have a more mixed distribution. Anchor-text items occur very frequently in some newspapers. An important reason for this, is that anchor-text items are typically small, making it possible to place a large number of these on a page. Another interesting observation is that URL-text-URL items are very common in Norwegian newspapers, but less frequent in international newspapers. This also shows how

[8] `http://dn.no/`

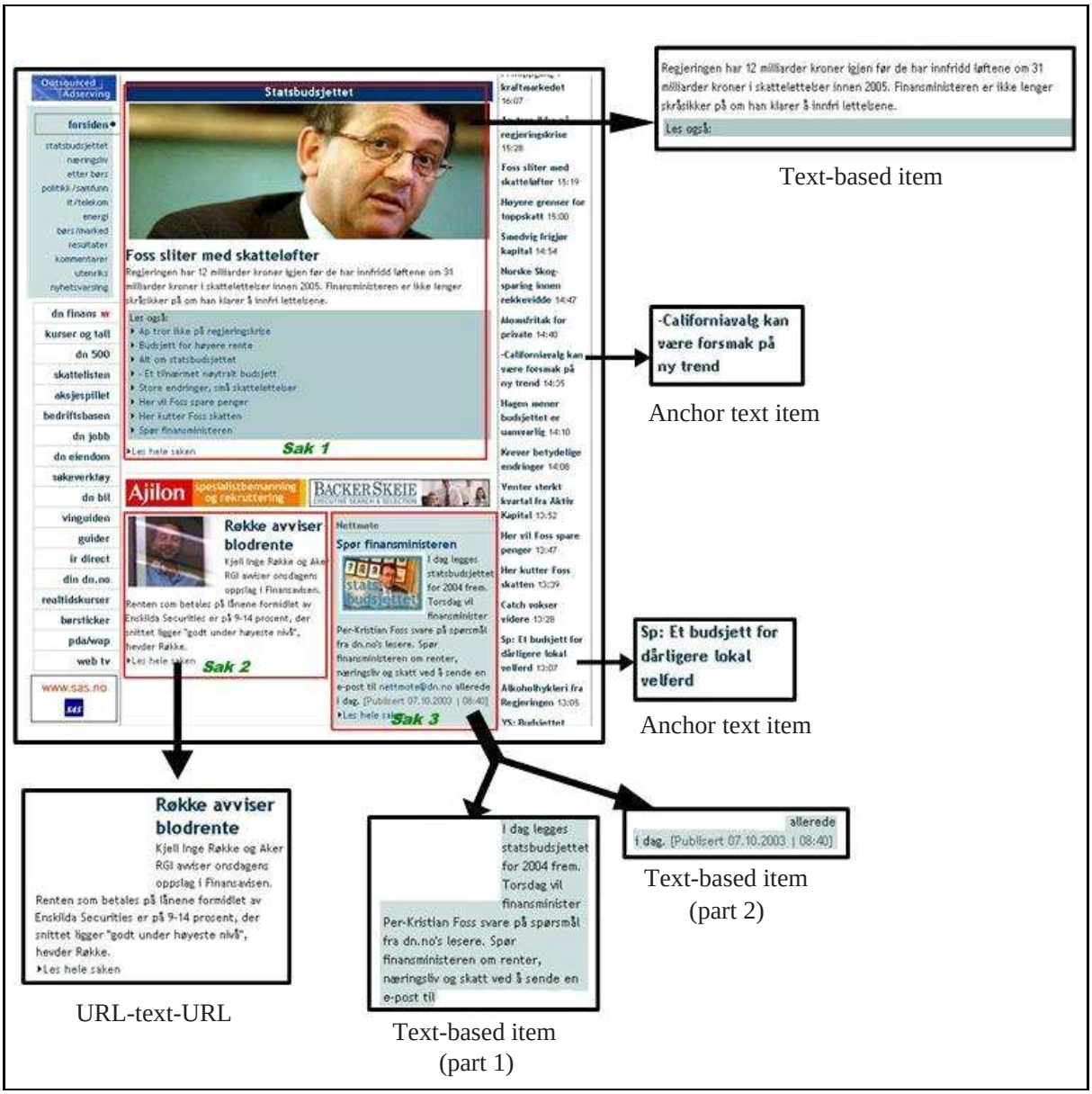


Figure 5. News items extracted from parts of dn.no**.**

Pattern	Norwegian	English
URL-text-URL	96.7%	86.9%
Anchor text	90.6%	93.6%
Text-based	79.4%	71.6%
Total	89.7%	84.0%

Table 1. Precision of news item extraction.

important it is in a study like ours to use a good mix of newspapers.

We also studied the number of news items on each page, illustrated on Figure 6(b). It shows an average of approximately 50 news items per page. The average length of news items was found to be in the order of 15 words. However, this number can be very dependent of the parameters used in the pattern-detection, by accepting items smaller than 5 words, the average size would also be lower.

4.2 Extraction quality

In order to quantify the extraction quality we can use the information retrieval measure *precision*, i.e., the number of relevant items detected (what a user would consider a news item when visually inspecting the page) as a fraction all the extracted items (all items output by the extraction tool). Table 1 gives a summary of the precision of news item extraction using our extraction tool. Table 2 gives a more detailed overview of precision for individual newspapers. It is obvious that URL-text-URL and anchor items are accurately detected, while detection of text-based items is less robust.

Precision can be also considered as an indicator of noise in the result. Another measure that can be used to judge the quality, is *recall*, i.e., the number of relevant items detected as a fraction of all news items on the page. One of our assumptions behind the extraction process is that given the fact that there will be mistakes in the automatic detection of news items, it is better with respect to the subsequent data mining process to have a few news items too much than potentially missing a few important items. The result is high recall, at the cost of a possibly reduced precision. However, as Table 2 shows, high recall has not significantly hurt precision in our case.

5 Mining news items

The motivation for the news item extraction is to increase the quality of the data mining by having more focused items to mine, i.e., news items instead of whole pages. In order to verify the applicability of our news item based approach to data mining, we have performed some experiments where we have studied data mining using the extracted news items. We will in this section report some results from clustering news items.

5.1 Text clustering

The goal of text clustering is to create groups of similar documents (or in our case, document fragments like news items), this means to automatically group document according to topic. This is useful in a number of contexts, for example:

- Given one news item, for example the result of a text query, find related news items.
- Hierarchical document clustering can be used to discover topic hierarchies in large document collections [12].

Clustering of high-dimensional data is difficult, making text clustering in particular a hard problem because of a very high number of dimensions. However, a number of dimensionality-reducing techniques exist, for example principal component analysis and singular value decomposition [6]. It is also possible to reduce the number of dimensions by removing stop words (see below), or only use the most interesting terms for clustering [16].

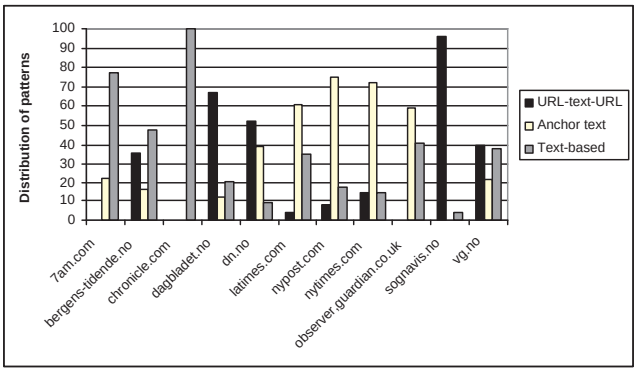
5.2 Data cleaning

In the system previous versions of web pages are kept. Even if a news item occurs in several versions of a page, it should only be stored once (in order to not reduce the quality of the data mining process). When a news item is detected it is compared with the news items in the previous version of the page,[9] and if both text *and* URL are identical the news item is discarded.

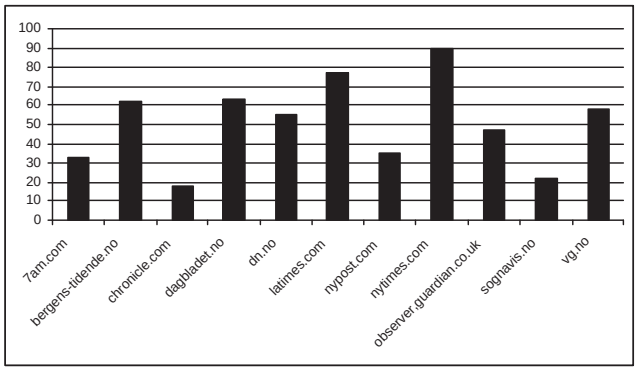
Some characters (and white space) can be written in alternative ways in HTML, one alternative is to use a letter from a particular character set, for example “å”, and another alternative is to write it on the form “å”. In order to have a uniform representation of these for the subsequent data mining process, they are translated to a common representation. Also formatting information is removed, so that only clean text is stored as a news item, together with the associated URL.

The occurrence of some very frequently occurring words that in itself does not carry information can reduce the quality of the data mining process. These *stop words*, like “a”, “it”, and “and”, are also removed during the cleaning process. Stop words can differ between different domains and languages, therefore it is made possible for the user of the system to specify additional stop words.

[9] Hash values are used in order to reduce the cost of comparing the news items.



(a) Distribution of extracted news items between different patterns (in %).



(b) Total number of news items extracted for each web page.

Figure 6. Web page statistics.

5.3 Clustering software

As a starting point in our study of news item clustering, we have used publically available software (described in more detail in [5]) to perform the clustering and visualization task:

- MC[10] takes as input a collection of text files, and creates vector-space models that can be used for text mining by Gmeans. MC uses the *compressed column storage technique* in order to reduce the space usage of the matrix.
- Gmeans[11] performs clustering of the vector/space model created by MC, using the K-means clustering algorithm.
- ClusterBrowser[12] generates a sequence of web pages that can be used to study the clustering results.

5.4 Results

In our experiments we have studied clustering of news items using the tools described above. Items from the Norwegian and English newspapers were studied separately. The size of the collections were approximately 100000 Norwegian news items and 44000 English news items. We studied the clustering using several number of clusters, from 50 to 1000 clusters in the result.

[10] Available from `http://www.cs.utexas.edu/users/jfan/dm/index.html`.

[11] Available from `http://www.cs.utexas.edu/users/dml/Software/gmeans.html`.

[12] Available from `http://www.cs.utexas.edu/users/yguan/datamining/clusterBrowser.html`.

The clustered news items are unlabeled, so quality measurements based on precision and recall is difficult because of the manual work it would involve. For the moment, we have instead tried to get an initial understanding of the quality by manual inspection of the news items in some of the clusters. Available information includes the number of items in each cluster, the identifiers/file names of the documents/items in a cluster, as well as a ranked list of the representative words characterizing each of the clusters (concept vector).

We will now give an example of the results of the clustering. For all cluster sizes when clustering Norwegian news items, one of the clusters had “hosein” as the highest ranked word in the concept vector. The concept vector for this cluster is shown in Figure 7. The background for many of the items in this cluster is essentially a Norwegian murder case, and the concept vector contains the name of the convicted murderer, and other words related to the murder and the following trial. Table 3 shows the highest ranked news items in this cluster when clustering news items into 1000 clusters, and these news items are strongly related to the murder case. Table 4 shows the lowest ranked news items, and the last of these items are part of another murder case, but still related in topic. The first of the news items in Table 4 also illustrates a case where the extraction has not completely succeeded: the end of the news item is actually the start of the next item.

6 Conclusions

Web newspapers provide a valuable resource for information, and text mining techniques can be applied to benefit more from the available information. However, because

Newspaper	URL-text-URL	Anchor-text	Text-based	Total
7am.com	-	62.5	100.0	90.9
bergens-tidende.no	95.5	100.0	93.3	95.2
chronicle.com	-	-	83.3	83.3
dagbladet.no	90.5	75.0	53.8	81.0
dn.no	100.0	95.2	40.0	92.6
latimes.com	100.0	97.8	70.4	88.2
nypost.com	66.7	96.2	33.3	82.9
nytimes.com	16.7	97.0	75.0	83.3
observer.guardian.co.uk	-	89.3	84.2	87.2
sognavis.no	100.0	-	0.0	95.5
vg.no	95.7	84.6	81.8	87.9

Table 2. Precision for individual newspapers (in %).

hosein (0.717),gamal (0.441), ars (0.21) , fengsel (0.207), domt (0.199), gry (0.175), ethvert (0.102), drapet (0.09), krever (0.082), griper (0.08), stotteordninger (0.08), ikkerhetsradet (0.079), norskmusikkgigantene (0.075), skien (0.074), kona (0.069), konsentrasjonsleirer (0.069), arild (0.067), agder (0.065), frps (0.062), soldaten (0.061), polsesnakk (0.058), dag (0.055), klarsignalfor (0.049), dommen (0.046), gripe (0.046), aktor (0.046), fredag (0.044), kone (0.043), dorum (0.036), betale (0.034), angripere (0.033), envoldelig (0.032), mottok (0.029), fjellheimen (0.026), spenningsleverandor (0.025), aksjepost (0.025), hagen (0.025), motiv (0.025), mener (0.023), hundesimulatoren (0.023), konen (0.022), juryen (0.022), nett (0.021), parets (0.021), garde (0.02), klappjakt (0.02), kamerat (0.02), veibommer (0.02), kjore (0.019), barn (0.019).

Figure 7. Top 50 words in concept vector (with associated ranks) for the "hosein" cluster.

Time/URL/ItemID and rank	Text
20031031_2010/www.p4.no /www.p4.no_case1(0.874)	Hosein fikk ni ar Gamal Hosein ble i dag domt til ni ars fengsel for drapet pa kona Gry Hosein. Les hele saken
20031101_2010/www.adressa.no /nyheter/innenriks/ www.adressa.no_nyheter _innenriks_case10(0.837)	Gamal Hosein domt til ni ars fengsel Gamal Hosein ble fredag domt til ni ars fengsel for drapet pa kona Gry Hosein. Han ma ogsa betale 250.000 kroner til hvert av parets to barn. Les hele saken

Table 3. Highest ranked news items for the "hosein" cluster.

Time/URL/ItemID and rank	Text
20031031_0410/ nyheter.no/innenriks /nyheter.no_innenriks _innenriks-20-1-1.html_case3 (0.119)	Telemarks Avis 31.10.2003 02:28 SKIEN: Liv Sofie Ostensen (52) tygger na pa et tilbud om a gi ut en bok om sine erfaringeri forbindelse med drapet pa datteren Gry. PublisertGAMAL DOMT IGJEN
20031106_1110/ www.bergens-tidende.no/ www.bergens-tidende.no _case23 (0.06)	06. nov, 06:00Drapssiktet husvert loslatt Gulating lagmannsrett mener det er sannsynlig at husverten (31) var med pa a planlegge drapet pa Mentz Dankert Lotvedt. Likevel loslot de KrF-eren i gar.

Table 4. Lowest ranked news items for the "hosein" cluster.

each newspaper page often contains news items of unrelated topics, page-based clustering will seldom give useful results. In order to improve on complete-page mining, we have in this paper presented an approach based on extracting the individual news items from the web pages and mining these separately. The extraction can be performed automatically using strategies based on pattern-detection strategies presented in this paper.

Based on a study of the quality of the news item extraction we confirm the quality and robustness of our strategies. In order to verify the applicability of our approach in text mining we have also performed clustering of the news items, and presented some results from these experiments.

Two additional aspects of news item extraction should also be mentioned: 1) the removal of non-relevant information also makes news item extraction useful as a way of data cleaning, and 2) the space needed for storing the news items is much smaller than what is needed for the pages themselves, typically only 10% of the original size, and this means that the news items can be used as summary information for web sites and stored in a web warehouse.

Future work includes a more detailed study of data mining techniques applied on the news items as well as a closer integration of the techniques into the V2 temporal document database system [14].

Acknowledgments

Part of this work was done when the first author visited Athens University of Economics and Business in 2004, supported by grant #145196/432 from the Norwegian Research Council.

References

[1] Z. Bar-Yossef and S. Rajagopalan. Template detection via data mining and its applications. In *Proceedings of the eleventh international conference on World Wide Web*, 2002.

[2] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the web. *WWW6/Computer Networks*, 29(7-13), 1997.

[3] D. Cai, S. Yu, J. Wen, and W. Ma. Extracting content structure for web pages based on visual representation. In *Web Technologies and Applications: 5th Asia-Pacific Web Conference (APWeb 2003)*, 2003.

[4] C. Chang, C. Hsu, and S. Lui. Automatic information extraction from semi-structured web pages by pattern discovery. *Decision Support Systems*, 35(1):129 – 147, April 2003.

[5] I. S. Dhillon, J. Fan, and Y. Guan. Efficient clustering of very large document collections. In *Data Mining for Scientific and Engineering Applications*. Kluwer Academic Publishers, 2001.

[6] C. Ding, X. He, H. Zha, and H. D. Simon. Adaptive dimension reduction for clustering high dimensional data. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)*, 2002.

[7] D. W. Embley, Y. Jiang, and Y.-K. Ng. Record-boundary discovery in web documents. In *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, 1999.

[8] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2000.

[9] IBM intelligent miner for text, `http://www-306.ibm.com/software/data/iminer/fortext/`.

[10] H. Kao, S. Lin, J. Ho, and M. Chen. Mining web informative structures and contents based on entropy analysis. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):41–55, 2004.

[11] 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.*, 31(2):84–93, 2002.

[12] B. Larsen and C. Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999.

[13] S. Lin and J. Ho. Discovering informative content blocks from web documents. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002.

[14] K. Nørvåg. V2: a database approach to temporal document management. In *Proceedings of the 7th International Database Engineering and Applications Symposium (IDEAS)*, 2003.

[15] L. Ramaswamy, A. Iyengar, L. Liu, and F. Douglis. Techniques for efficient fragment detection in web pages. In *Proceedings of the twelfth international conference on information and knowledge management*, 2003.

[16] S. M. Rüger and S. E. Gauch. Feature reduction for document clustering and classification. Technical Report 8, Department of Computing. Imperial College of Science, Technology and Medicine, 2000.

[17] A. Sahuguet and F. Azavant. Building intelligent web applications using lightweight wrappers. *Data Knowl. Eng.*, 36(3):283–316, 2001.

[18] L. Yi, B. Liu, and X. Li. Eliminating noisy information in web pages for data mining. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003.