Journal of Web Engineering, Vol. 8, No. 3 (2009) 268–284
© Rinton Press

AN AUTOMATIC WEB NEWS ARTICLE CONTENTS EXTRACTION SYSTEM BASED ON RSS FEEDS

HAO HAN TOMOYA NORO TAKEHIRO TOKUDA
Department of Computer Science, Tokyo Institute of Technology
Ookayama 2-12-1-W8-71, Meguro, Tokyo 152-8552, Japan
{*han, noro, tokuda*}*@tt.cs.titech.ac.jp*

Received December 17, 2008
Revised September 2, 2009

Nowadays, the Web news article contents extraction is vital to provide news indexing and searching services. Most of the traditional methods need to analyze the layout of news pages to generate the wrappers manually or automatically. It is a costly work and needs much maintenance during the extraction over a long period of time. In this paper, we construct an automatic Web news article contents extraction system based on RSS feeds. We propose an effective and efficient algorithm to extract the news article contents from the news pages without the analysis of news sites before extraction. We calculate the relevance between the news title and each sentence in the news page to detect the news article contents. Our approach is applicable to the general types of news RSS feeds and independent of news page layout. Our experimental results show that our approach can extract the news article contents automatically, accurately and constantly.

Keywords: Web News Article, Information Extraction, RSS Feed

Communicated by: B. White & R. Baeza-Yates

1 Introduction

The Internet has marked this era as the information age. The Web is rapidly moving towards a platform for mass collaboration in content production and consumption, and the increasing number of people are turning to online source for their daily news. Traditional newspapers have developed significant Web presences. Fresh contents on a variety of topics, people, and places are being created and made available on the Web at a breathtaking speed. We can extract and analyze the Web news article contents to acquire the desired information and knowledge.

Our purpose is to extract the Web news article contents from news sites over a long period of time. Usually, the extraction process includes two steps. Firstly, the news sites are crawled to collect the news pages. Secondly, the news article contents are extracted from news pages.

The news sites comprise different kinds of Web pages. Besides the news pages, there are many non-news pages, such as the blog, shopping, weather, advertisement, yellow pages and even same pages with different URLs. Furthermore, these news pages are spread in the different sections of news sites. The news sites are crawled to find as many news pages as possible, but actually, it is difficult to recognize and acquire all the news pages quickly from a large number of Web pages.

Usually, besides the news title and news contents, there are other elements on a news page such as the advertisement, related stories and comments. In order to recognize and extract the parts of news article contents from the full text of news pages, wrappers are generated based on the analysis of layout of news pages by many extraction methods. Web page layout is the style of graphic design in which text or pictures are set out on a Web page. The different news sites use the different news page layout, and each news site uses more than one layout. Therefore, most extraction methods analyze the news page layout of each news site for news contents extraction. It is a costly work.

Moreover, the news sites update the layout of news pages irregularly. If the news sites update the layout of news pages, the corresponding analysis has to be done again.

Overall, it is not easy to extract the news article contents from news sites efficiently and quickly over a long period of time for the traditional methods. They have to face these two problems: "*how to collect the news pages from news sites?*" and "*how to extract the news article contents correctly if news sites update the layout of news pages?*".

In this paper, we propose an approach to construct an automatic Web news article contents extraction system based on RSS feeds. RSS is a family of Web feed formats used to publish frequently updated content such as news headlines. We can collect the latest news pages from news RSS feeds conveniently as soon as they are published. We give an effective algorithm to extract the news article contents from the news pages automatically. Instead of the methods like pattern matching and machine learning, our extraction method is independent of Web page layout and does not need to analyze the news sites before extraction. We calculate the relevance between the news title and each sentence to detect the news paragraphs from the full text of the news page. Our approach is applicable to the general news pages from the general news RSS feeds, and realizes the news article contents extraction over a long period of time without the maintenance task.

In the next section we present the motivation of our research and an overview of the related work. We explain our news article contents extraction approach in detail in Section 3. In Section 4, we give an evaluation of our system and explain the implementation of our system. Finally, we discuss the future work and give the conclusion in Section 5.

2 Motivation and Related Work

Taking advantage of the fact that there are a great and increasing number of news Web sites, a lot of approaches have been proposed for extracting the news article contents from the news Web sites.

Some approaches use the manual or semi-automatic example-based wrapper learning to extract the news article contents from the general news pages ultimately. Shinnou et al. gave an extraction wrapper learning method and expected to learn the extraction rules which could be applied to news pages from other various news sites [18]. Reis et al. gave a calculation of the edit distance between two given trees for the automatic Web news article contents extraction [5]. Fukumoto et al. gave the focus on subject shift and presented a method for extracting key paragraphs from documents that discuss the same event [7]. It uses the results of event tracking which starts from a few sample documents and finds all subsequent documents. However, if a news site uses too many different layouts in the news pages, the learning procedure costs too much time and the precision becomes low.

Zheng et al. presented a news page as a visual block tree and derived a composite visual feature set by extracting a series of visual features, then generated the wrapper for a news site by machine learning [23]. However, it uses manually labeled data for training and the extraction result may be inaccurate if the training set is not large enough. The similar problems may occur to [4] and [24].

Webstemmer [19] is a Web crawler and HTML layout analyzer that automatically extracts main text of a news site without having banners, advertisements and navigation links mixed up. It analyzes the layout of each page in a certain web site and figures out where the main text is located. All the analysis can be done in a fully automatic manner with little human intervention. However, this approach runs slowly at contents parsing and extraction, and sometimes news titles are missing. TSReC [10] provides a hybrid method for news article contents extraction. It uses tag sequence and tree matching to detect the parts of news article contents from a target news site. However, for these methods, if the news sites change the layout of news pages, the analysis of layout or tag sequence has to be done again.

Some approaches analyze the features of news pages to generate the wrappers for automatic or semi-automatic extraction. CoreEx [16] scores every node based on the amount of text, the number of links and additional heuristics to detect the news article contents. However, it does not seem to deal with the news pages with many related information in text format and may miss the title information when the news article header appears far away from the body. Dong et al. gave a generic Web news article contents extraction approach based on a set of predefined tags [6]. The experiment of this method is based on the assumption that the news pages from a news site use the same layout. But, actually, there are many different layouts used in a news site.

There are some layout-independent extraction approaches. TidyRead [20] and Readability [17] render Web pages with better readability as an-easy-to-read manner by extracting the context text and removing the cluttered materials. They run as plug-in or bookmarklet of Web browser. However, the extraction result is a part of Web page containing the HTML tags. It also contains some other non-news elements such as the advertisement and related links. Parapar et al. gave a set of heuristics to identify and extract the news articles [15]. It is an effective method and the extraction process is independent of the layout of each news page. Wang et al. proposed a wrapper to realize the news articles extraction by using a very small number of training pages based on machine learning processes from news sites [21]. The algorithm is based on the calulation of the rectangle sizes and word numbers of news title and contents. However, these approaches still need to set the values of some parameters manually, and could not be proved to extract the news article successfully or automatically if news sites update the layout of news pages. Full-Text RSS [8] only returns the news article contents when the supplied RSS feed has a summary or description of some kind.

These approaches are still not widely used, mostly because of the need for high human intervention or maintenance, and the low quality of the extraction results. Most of them have to analyze the news pages from a news site before they extract the news article contents from this news site. To address these problems, we propose a new approach to realize the automatic news article contents extraction based on RSS feeds. Compared with the developed work, our approach has the following features.

- We can collect the news pages from the news RSS feeds easily because more and more news sites distribute the latest news by RSS feeds. For example, a study shows that 97% of America's top newspaper websites provide RSS feeds [1].
- Our news article contents extraction algorithm is applicable to the general types of news pages. We do not need the methods like pattern matching or machine learning, which cost too much time before the extraction. We just give the RSS feeds to our extraction system and all the extraction can be completed automatically. Our experimental results show that our extraction system has a high extraction precision over a long period of time.
- Our extraction system is constructed easily and does not need any maintenance during the long period extraction.

We explain our approach and give an evaluation in the following sections.

3 News Article Contents Extraction

As shown in Fig. 1, we extract the news title and URL of each news page from RSS feeds. We split the news title to get the keyword list and use them to detect the position of news title in the news page. Then, we recognize one paragraph of news article by news title position and keyword list. Finally, we find all the paragraphs of news article contents and extract them out of the full text of the news page. We explain our algorithm step by step in this section.

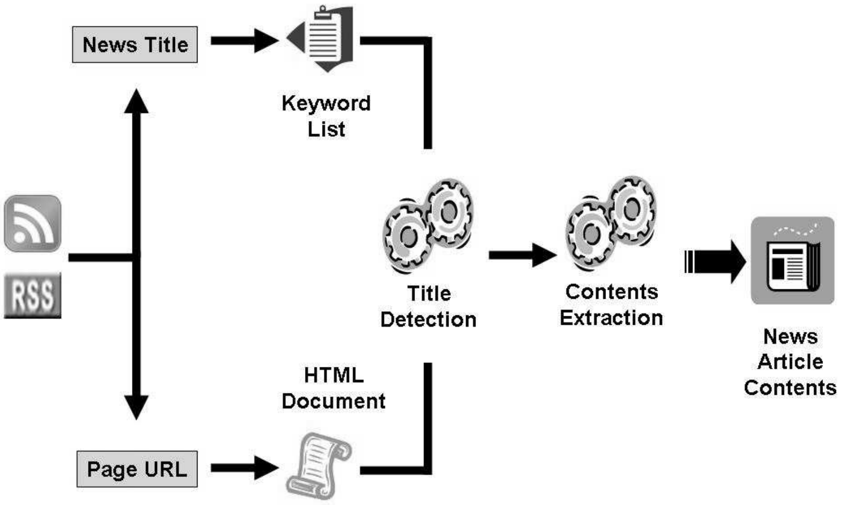


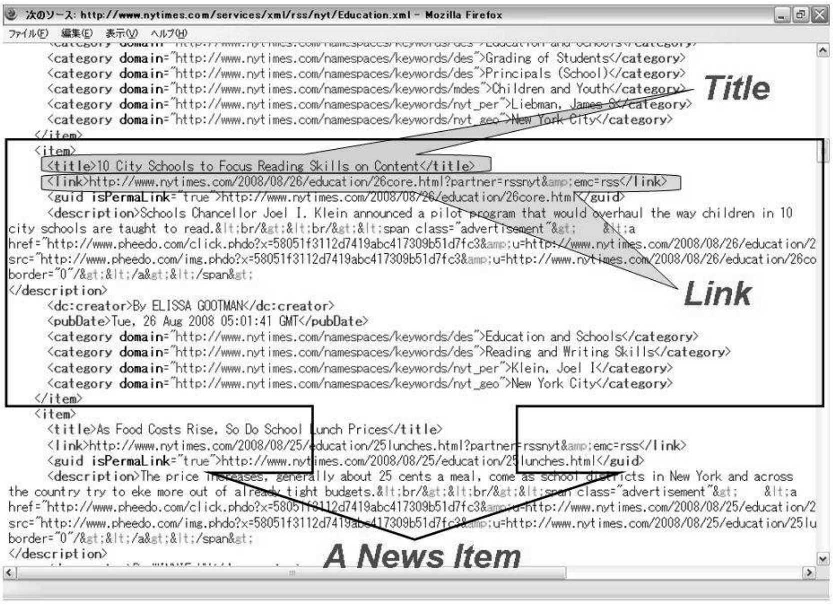
Fig. 1. The outline of RSS news article contents extraction system

3.1 News RSS Feeds

There are more and more Web sites distributing news by the news RSS feeds for the easy access and subscription of news. RSS is an XML-based format for sharing and distributing frequently updated Web contents such as news and blogs. A news RSS document, which is called a "news RSS feed", usually contains headlines, summaries and links to news pages. RSS makes it possible for people to keep up with their favorite Web sites in an automated manner that's easier than checking them manually. It broadcasts in the following standardized formats.

- Really Simple Syndication (RSS 2.0)
- RDF Site Summary (RSS 1.0 and RSS 0.90)
- Rich Site Summary (RSS 0.91)

Although there are a number of different formats of RSS, all of them include the link and title information in <link> and <title> respectively. These two information fields are the minimum necessary parts of each news item in a RSS feed as shown in Fig. 2.


次のソース: http://www.nytimes.com/services/xml/rss/nyt/Education.xml - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) ヘルプ(H)
Title
Link
A News Item

```
      <category domain="http://www.nytimes.com/namespaces/keywords/des">Grading of Students</category>
      <category domain="http://www.nytimes.com/namespaces/keywords/des">Principals (School)</category>
      <category domain="http://www.nytimes.com/namespaces/keywords/mdes">Children and Youth</category>
      <category domain="http://www.nytimes.com/namespaces/keywords/nyt_per">Liebman, James S</category>
      <category domain="http://www.nytimes.com/namespaces/keywords/nyt_geo">New York City</category>
    </item>
    <item>
      <title>10 City Schools to Focus Reading Skills on Content</title>
      <link>http://www.nytimes.com/2008/08/26/education/26core.html?partner=rssnyt&amp;emc=rss</link>
      <guid isPermaLink="true">http://www.nytimes.com/2008/08/26/education/26core.html</guid>
      <description>Schools Chancellor Joel I. Klein announced a pilot program that would overhaul the way children in 10
city schools are taught to read.&lt;br/&gt;&lt;br/&gt;&lt;span class="advertisement"&gt;    &lt;a
href="http://www.pheedo.com/click.phdo?x=58051f3112d7419abc417309b51d7fc3&amp;u=http://www.nytimes.com/2008/08/26/education/2
src="http://www.pheedo.com/img.phdo?x=58051f3112d7419abc417309b51d7fc3&amp;u=http://www.nytimes.com/2008/08/26/education/26co
border="0"/&gt;&lt;/a&gt;&lt;/span&gt;
</description>
      <dc:creator>By ELISSA GOOTMAN</dc:creator>
      <pubDate>Tue, 26 Aug 2008 05:01:41 GMT</pubDate>
      <category domain="http://www.nytimes.com/namespaces/keywords/des">Education and Schools</category>
      <category domain="http://www.nytimes.com/namespaces/keywords/des">Reading and Writing Skills</category>
      <category domain="http://www.nytimes.com/namespaces/keywords/nyt_per">Klein, Joel I</category>
      <category domain="http://www.nytimes.com/namespaces/keywords/nyt_geo">New York City</category>
    </item>
    <item>
      <title>As Food Costs Rise, So Do School Lunch Prices</title>
      <link>http://www.nytimes.com/2008/08/25/education/25lunches.html?partner=rssnyt&amp;emc=rss</link>
      <guid isPermaLink="true">http://www.nytimes.com/2008/08/25/education/25lunches.html</guid>
      <description>The price increases, generally about 25 cents a meal, come as school districts in New York and across
the country try to eke more out of already tight budgets.&lt;br/&gt;&lt;br/&gt;&lt;span class="advertisement"&gt;    &lt;a
href="http://www.pheedo.com/click.phdo?x=58051f3112d7419abc417309b51d7fc3&amp;u=http://www.nytimes.com/2008/08/25/education/2
src="http://www.pheedo.com/img.phdo?x=58051f3112d7419abc417309b51d7fc3&amp;u=http://www.nytimes.com/2008/08/25/education/25lu
border="0"/&gt;&lt;/a&gt;&lt;/span&gt;
</description>
```

Fig. 2. A RSS feed of New York Times

Fig. 3 shows a simple example of RSS feed. We parse the RSS feed to extract the node values of <link> and <title>, which are the link to news page and the title of news respectively. We use the link to extract the HTML document of each news page from the news site, and use the title information to complete the news contents extraction in the following algorithm description.

```
<?xml version="1.0"?>
<rss version="2.0">
  <channel>
    <title>Example</title>
    <link>http://www.example.com/example.rss</link>
    <description>This is an example RSS feed.</description>
    <language>en-us</language>
        ...
    <item>
      <title>A Monkey Stopped Morning Commuters at Shibuya - One
      of Tokyo's Busiest Subway Stations</title>
      <link>http://www.example.com/news/example1.html</link>
    </item>
    <item>
      <title>This is the title of another example news</title>
      <link>http://www.example.com/news/example2.html</link>
    </item>
    <item>
      ...
    </item>
    ...
  </channel>
</rss>
```

Fig. 3. A simple example of news RSS feed

3.2 Title Keywords Acquisition

The news title is a piece of important information for the recognition of the news contents from the full text of news page. If we correctly locate the position of the title in a news page, the position of news contents text would be found easily because the contents text is a list of paragraphs closely preceded by the title usually as shown in Fig. 4. In addition, for a news article, the contents describe the same topic of news title in detail, and the words constituting the title would occur in the news contents frequently usually. We split the title sentence into single words to make a list of keywords and use these keywords to find out the news article contents from the news page.

We create a keyword list K from the news title using the following steps.

1. We split the news title sentence into a word list K using whitespace as the delimiter.

2. We prepare a stop word list including the articles, prepositions and conjunctions. We remove these stop words from the word list K.

3. We remove the characters ” *’s*” or ” *s’*” from the words ending with ” *’s*” or ” *s’*” in list K. For example, we replace ” *Tom’s*” with ” *Tom*”, and replace ”*parents’*” with ”*parent*”.

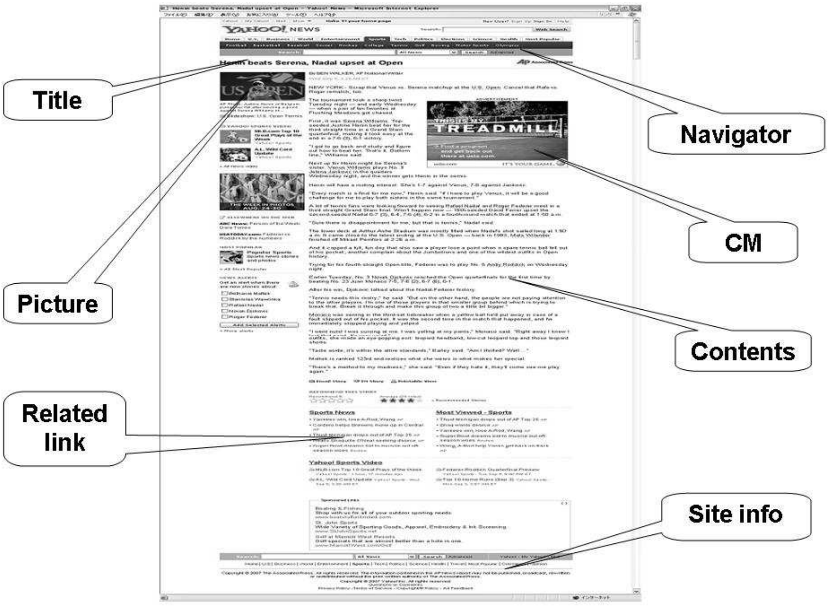


Fig. 4. News contents text is a list of paragraphs closely preceded by the title usually

The words in list K are the final keywords. The following is an example of the keywords acquisition.

”A Monkey Stopped Morning Commuters at Shibuya - One of Tokyo’s Busiest Subway Stations”

$\Downarrow$

”Monkey”, ”Stopped”, ”Morning”, ”Commuters”, ”Shibuya”, ”One”, ”Tokyo”, ”Busiest”, ”Subway”, and ”Stations”

3.3 Full Text Analysis

An HTML document may be represented as a tree structure. A sentence in a Web page is a visible character string, which is the value of a leaf node. It is possible for each sentence to be the title or a paragraph. We use the following steps to analyze the full text of a news page in order to find the most possible title and paragraphs.

1. For each leaf node N, create a word list W_N from the text in the same way as the keyword acquisition method described in Section 3.2.

2. For each node N, calculate $WordNumber_N$ and $KeyNumber_N$ as follows.

$$WordNumber_N = \begin{cases} |W_N| & \text{if } N \text{ is a leaf node} \\ \sum_{n \in \mathrm{Child}_N} WordNumber_n & \text{otherwise} \end{cases}$$

$$KeyNumber_N = \begin{cases} |W_N \cap K| & \text{if } N \text{ is a leaf node} \\ \sum_{n \in \mathrm{Child}_N} KeyNumber_n & \text{otherwise} \end{cases}$$

Where Child_N is a set of child nodes of the node N. Fig. 5 shows an example of full text analysis.

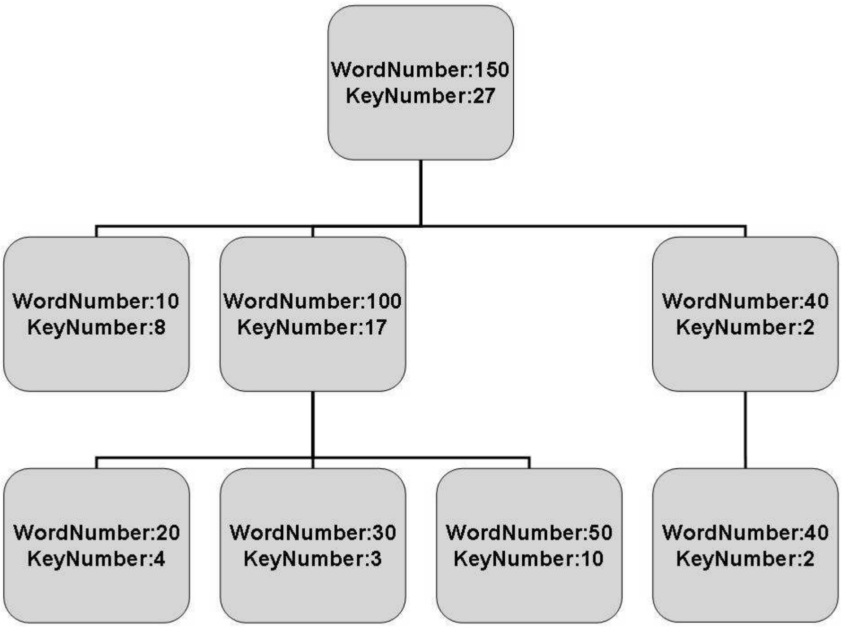


Fig. 5. A full text analysis example

3.4 News Title Detection

After the full text analysis, we need to find out the real news title in news page using the news title extracted from the news RSS feed. Usually, the real news title in a news page is same or similar to the news title in news RSS feed. We use the following formula to calculate the similarity between each sentence of the news page and the news title in news RSS feed.

$$Similarity = \frac{KeyNumber^2}{WordNumber \times TitleKeywordNumber}$$

Where, $KeyNumber$ and $WordNumber$ are the attribute of the corresponding node of each sentence respectively, and $TitleKeywordNumber$ is the size of keywords list of news title in news RSS feed.

We think a sentence is a possible real news title in the news page if the value of $Similarity$ is more than a predetermined threshold, and a node whose value is a possible news title would be a possible title node. Assuming that $WordNumber$ and $KeyNumber$ of each node are calculated as shown in Fig. 5 and TitleKeywordNumber is 8, Similarity of each leaf node is calculated as shown in Fig. 6. Assuming that the predetermined threshold is set to 0.6, the node A is judged as the title node.

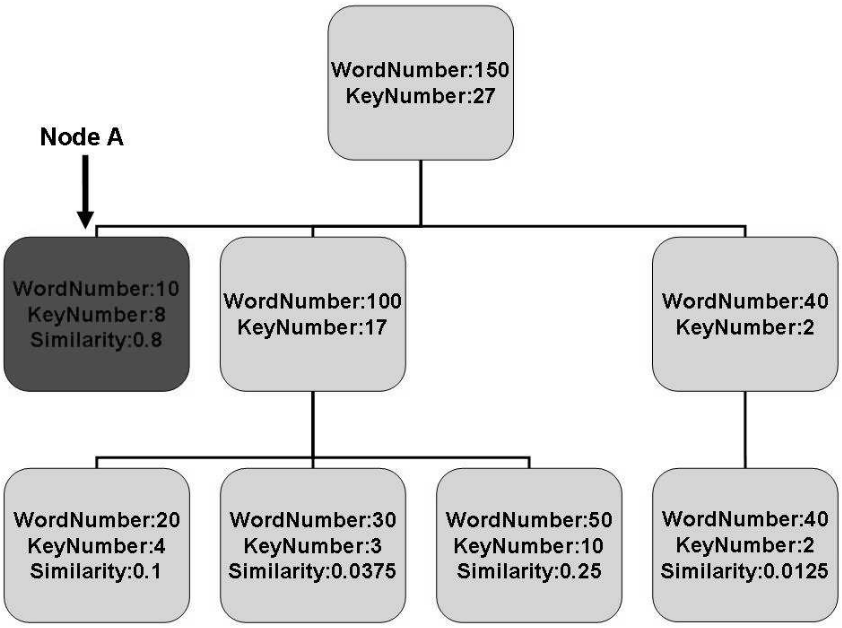


Fig. 6. A news title detection example

However, the news title in news RSS feed is not always same or similar to the real news title in the corresponding news page. In some news sites, we even can find that the news title in news RSS feed is different from the real news title totally, but same as the other sentences in the news page such as the related links like shown in Fig. 7. Moreover, some news titles are so short and simple that we can find two or more same or similar sentences in news pages. Therefore, there are five different situations about the possible news title and the real title in a news page.

1. There is no possible news title.

2. There is just one possible news title, and it is the real news title in the news page.

3. There is just one possible news title, but it is not the real news title in the news page.

4. There are two or more possible news titles, and one of them is the real news title in the news page.

5. There are two or more possible news titles, but none of them is the real news title in the news page.

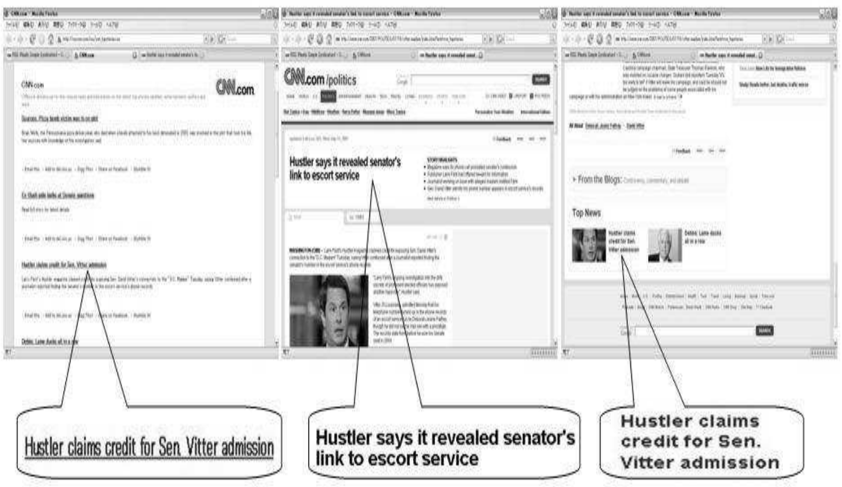

Hustler claims credit for Sen. Vitter admission
Hustler says it revealed senator's link to escort service
Hustler claims credit for Sen. Vitter admission

Fig. 7. A news title in a news RSS feed of CNN is different from the real news title, but same as a related link in the news page

3.5 News Paragraph Recognition

Usually, the news contents text is a list of paragraphs immediately below the title. It becomes easier to recognize the paragraphs from the full text after we find the real news title in the news page. However, we can not always guarantee the found possible news title is the real news title in the news page as we describe in Section 3.4. In our method, we divide an HTML document into one or more *contents ranges* and at most one *reserve range* as follows.

1. If no node is judged as the possible title node, the whole part of the document is a contents range (It has no reserve range) (Fig. 8(a)).
2. If only one node is judged as the possible title node, the document is divided into one contents range and one reserve range: the following part of the possible title node is a contents range and the preceding part is a reserve range (Fig. 8(b)).
3. If two or more nodes are judged as the possible title nodes, the document is divided into some contents ranges and a reserve range: the part between each possible title node (except the last possible title node) and the next possible title node is a contents range, the following part of the last possible title node is a contents range, and the preceding part of the first possible title node is a reserve range (Fig. 8(c)).

Then, we try to find a news paragraph in the contents ranges by calculating *Possibility* (described later). If a news paragraph cannot be found and the document has a reserve range, we search for the reserve range.

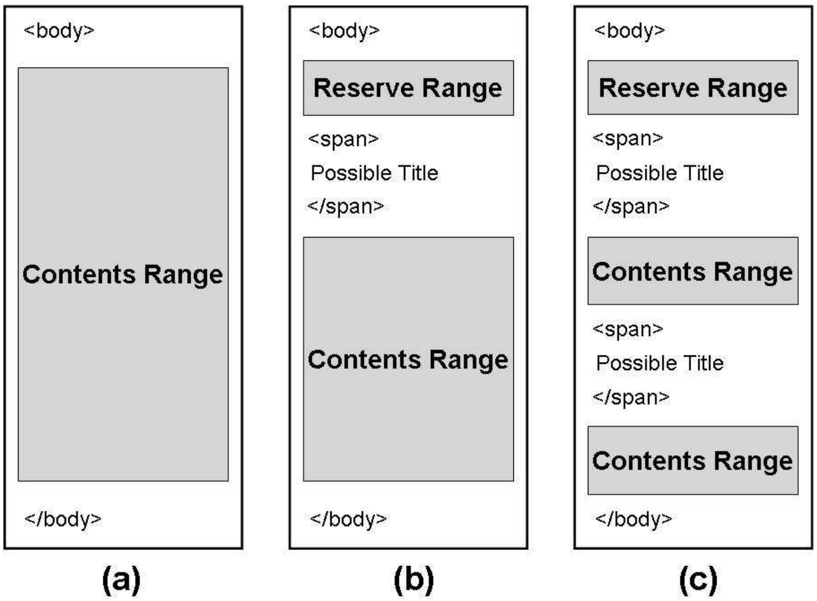


Fig. 8. Contents range and reserve range

A node whose value is a paragraph is a paragraph node. Although each node of each selective range, including the link text node, is possible to be one of the paragraph nodes, most of the paragraphs of a news article are non-link texts. We give a possibility value for each non-link node and select one with the highest possibility value as the final most possible paragraph node if the highest possibility value is more than a predetermined threshold. If the highest possibility value is less than this predetermined threshold, we would find a node with the highest possibility value in the reserve range and then compare the two possibilities to select one with the higher possibility value as the final most possible paragraph node. We use the following formula to calculate the possibility value of each non-link node.

$$Possibility = WordSum \times (KeySum + 1)$$

Where, $WordSum$ is the sum of the attributes $WordNumber$ of each node and its related nodes. $KeySum$ is the sum of the attributes $KeyNumber$ of each node and its related nodes. For Node A and Node B, we think Node B is a related node of Node A if Node B satisfies the following conditions.

1. Node B and Node A are in the same selective range.

2. Node B and Node A are sibling nodes, or their parent nodes are sibling nodes.

3. Node B or its parent node is of one of the following nodes: $\sharp text$, <span>, <a>, <p>, <ul>, <ol>, <dd>, <dt>, <strong>, <h1>, <h2>, <h3>, <h4>, and <b>.

Assuming that $WordNumber$ and $KeyNumber$ of each node are calculated as shown in Fig. 5 and node A is judged as the title node as shown in Fig. 6, node B, C, D and E belong to the contents range as shown in Fig. 9. Assuming that the predetermined threshold is set to 100, the node B is judged as the paragraph node.

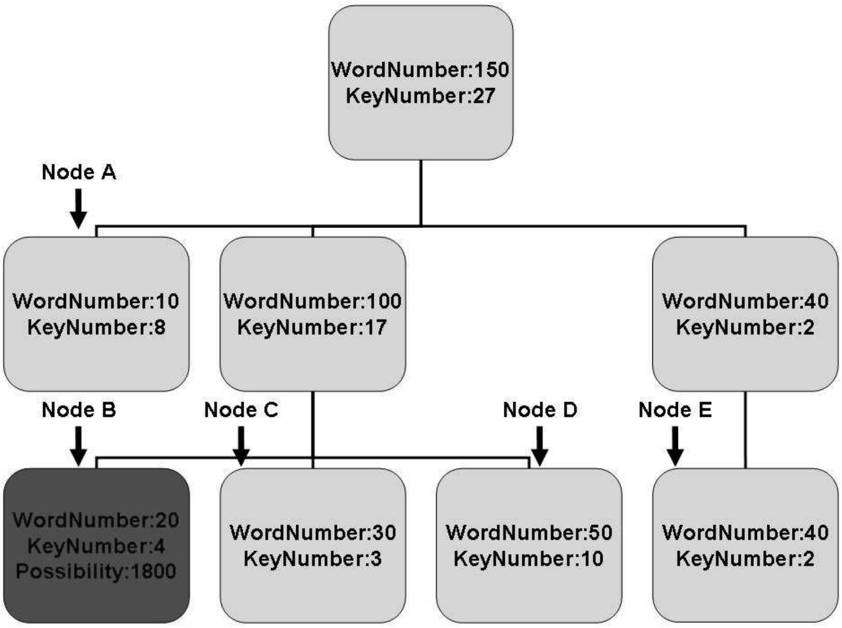


Fig. 9. A paragraph recognition example

3.6 News Contents Extraction

After the paragraph recognition, we get a paragraph of the entire news contents text. Usually, the entire contents text of news is a list of continuous paragraphs. However, there is advertisement information such as the image advertising among the paragraphs of news in some news sites. We need to find out the text nodes representing the news contents. If Node A is a node whose node value is a paragraph, we think Node B is also a paragraph if Node B satisfies the following conditions.

1. Node B and Node A are sibling nodes, or their parent nodes are sibling nodes.
2. Node B or its parent node is of one of the following nodes: $\sharp text$, <span>, <a>, <p>, <ul>, <ol>, <dd>, <dt>, <strong>, <h1>, <h2>, <h3>, <h4>, and <b>.

Finally, we get a list of nodes of which each one represents a paragraph of the news contents text like shown in Fig. 10. However, in some news pages, the title node is a sibling node of paragraph nodes and has the same node name. It uses a (different) class attribute and shows the title in a different style such as the font and color. Here, we pick out the title node from the node list (usually the first one) and the rest nodes are the paragraph nodes. We get the node value, which is text information, from each paragraph node as a paragraph. The extracted news contents text is the extracted continuous text information.

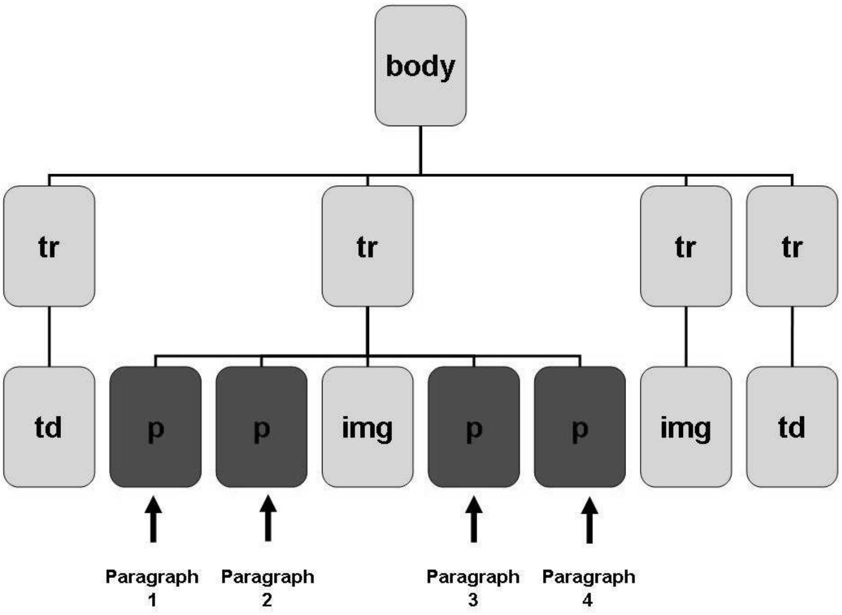


Fig. 10. A news article full contents extraction example

4 Evaluation and Implementation

In this section, we present the evaluation of our approach and explain the implementation of our extraction system. After we analyzed a large number of news pages from many news sites, we set the threshold as 0.6 in Section 3.4 and 100 in Section 3.5 respectively based on the statistical results. In the following experimental results, we prove that the thresholds are suitable for the general news pages.

We constructed a news article contents extraction system and collect the news from 641 RSS feeds distributed by 38 different news Web sites listed in Table 1. These selected news Web sites are the most popular on-line news publishers, including the global and domestic news sites from 19 countries/regions. We check the RSS feeds once every four hours and extract the newly published news articles. Since May 2007, we have collected more than 1.8 million pieces of news articles.

Our experiments were run periodically using the news articles randomly collected from the extraction results. Our experimental results are listed in Table 2. Here, *Success* means that our extraction system extracts the news article contents correctly, and *Failure* means that our extraction system extracts nothing or partial paragraphs or other non-news parts such as advertisements and related stories.

Although the news sites update the layout of news pages irregularly, our news article

Table 1. A list of RSS news sites which we collected news articles from

Country/region	News site	URL
United States	CNN	http://www.cnn.com/
	ABC News	http://www.abcnews.go.com/
	New York Times	http://www.nytimes.com/
	Washington Post	http://www.washingtonpost.com/
	Wall Street Journal	http://online.wsj.com/
	Chicago Tribune	http://www.chicagotribune.com/
	USA Today	http://www.usatoday.com/
	United Press International	http://www.upi.com/
	Los Angeles Times	http://www.latimes.com/
	CNet	http://www.cnet.com/
	Fox News	http://www.foxnews.com/
	CBS	http://www.cbsnews.com/
	The Associated Press	http://www.ap.org/
	International Herald Tribune	http://www.iht.com/
United Kingdom	BBC	http://www.bbc.co.uk/
	Guardian Unlimited	http://www.guardian.co.uk/
	Reuters	http://www.reuters.com/
Canada	CNW Group	http://www.newswire.ca/
	CBC	http://www.cbc.ca/
China	Xinhuanet	http://news.xinhuanet.com/english/
	China.org.cn	http://www.china.org.cn/
France	Euro News	http://www.euronews.net/
	France 24	http://www.france24.com/france24Public/en/
Japan	Mainichi Daily News	http://mdn.mainichi.jp/
	Japan News Net	http://www.thejapannews.net/
Africa	All Africa	http://allafrica.com/
Australia	News.com.au	http://www.news.com.au/
Germany	Deutsche Welle	http://www.dw-world.de/
Holland	DutchNews	http://www.dutchnews.nl/
India	Hindustan Times	http://www.hindustantimes.com/
Indonesia	Jakarta News Net	http://www.jakartanews.net/
Italy	Agenzia Nazionale Stampa Associata	http://www.ansa.it/
Middle East	Al Jazeera	http://english.aljazeera.net/
Pakistan	The News International	http://www.thenews.com.pk/
Russia	RIA Novosti	http://en.rian.ru/
Singapore	The Straits Times	http://www.straitstimes.com/
South Korea	Chosun Ilbo	http://english.chosun.com/
Tailand	Bangkok Post	http://www.bangkokpost.com/

Table 2. The experimental results of extraction accuracy rate

Period	Sum	Success	Failure	Precision
May 2007 - Aug 2007	1000	970	30	97.0%
Sep 2007 - Jan 2008	500	491	9	98.2%
Feb 2008 - May 2008	500	485	15	97.0%
Jun 2008 - Sep 2008	500	488	12	97.6%
Total	2500	2434	66	97.4%

contents extraction method works well during each period from our experimental results and the precision of extraction is over 97%. The experimental results prove that our extraction algorithm is highly accurate over a long period of time. These periodical experiments are necessary and important for testifying that extraction algorithm is fully independent of the updated layout.

Besides the RSS feeds provided by the various news sites, our extraction method can also extract the news article contents from the customized RSS feeds of news aggregation sites such as the Google News [9], or the search results feeds of news site databases such as BBC news search engine by a metasearch method [22, 13, 2]. We have proved that our extraction method is suitable for the extraction from the these two types of RSS feeds quickly and accurately in [14, 12].

We implemented our proposed news article contents extraction method and the experimental results prove that our extraction algorithm is highly accurate. However, in some news pages, a paragraph, usually the outline of news article, shows in different style compared to other paragraphs. This kind of paragraph looks like a non-news part such as an advertisement in text format, and is omitted in the extraction. Moreover, some news article contents are too short to recognize from the news pages. For example, a news flash about baseball game result, which contains just a short paragraph of ten words, maybe can not be extracted correctly.

Compared with other developed extraction systems, our extraction system has the following strong points.

1. Our extraction system keeps a high extraction precision over a long period of time. Our periodical and continual experiments last more than seventeen months. The experimental results are convincible, which are not found in other related work [18, 5, 7, 23, 4, 24, 19, 10, 16, 6, 15, 21].

2. Our extraction system is constructed easily and does not need any maintenance during the long period extraction. We do not need to analyze the layout of news pages since our extraction algorithm is independent of the layout of Web pages. It does not need to reconfigure extraction even though the news sites update the layout of news pages.

3. Our extraction system runs quickly because our algorithms are simple and efficient. For the extraction of a large number of news articles, a simple algorithm of low computational complexity saves a considerable amount of time. Table 3 shows a list of costed time of each news article extraction. The time is the calculation cost at our local computer (CPU: Intel Xeon 2.66 GHz, Memory: 2 GB, OS: Vine Linux 3.3.2, JDK: 1.6.0) excluding getting news pages from news sites and saving the extraction results into database.

We implemented our extraction system to construct other news systems. We have constructed a news index system which supports users who would like to observe difference in various topics (e.g. economy, sports, health, education and culture) among countries/regions based on our news article contents extraction system [14]. We also have presented automatic methods for constructing news directory systems which contain the flat or hierarchical classification information of our extracted news article contents [11], and the personal news RSS feeds generation [12]. Because our method is independent of news page layout and applicable to the general news pages, it is more suitable to be integrated into other news related applications. For example, if it is integrated into News RSS Reader, the news article full-text retrieval could be realized.

Our extraction approach is not limited to the English news sites. By using the morphological analysis tools like ChaSen [3], we also could extract the news article contents from Japanese news sites.

Table 3. A list of costed time of each news article (milliseconds)

News site	Time	News site	Time
CNN	6.02	ABC News	2.72
New York Times	0.63	Washington Post	6.62
Wall Street Journal	6.80	Chicago Tribune	8.93
USA Today	4.25	United Press International	3.20
Los Angeles Times	3.70	CNet	7.20
Fox News	3.16	CBS	7.24
The Associated Press	2.08	International Herald Tribune	2.80
BBC	3.19	Guardian Unlimited	5.12
Reuters	4.09	All Africa	2.98
CNW Group	2.25	CBC	3.41
Xinhuanet	2.94	China.org.cn	3.23
Euro News	1.80	France 24	3.67
Mainichi Daily News	2.40	Japan News Net	2.99
News.com.au	5.12	Deutsche Welle	2.67
DutchNews	1.00	Hindustan Times	7.57
Jakarta News Net	2.21	Agenzia Nazionale Stampa Associata	3.50
Al Jazeera	2.75	The News International	2.94
RIA Novosti	1.67	The Straits Times	2.25
Chosun Ilbo	0.52	Bangkok Post	2.40

5 Conclusion

In this paper, we have presented an effective approach to realize the automatic news article contents extraction using the news RSS feeds. We proposed an algorithm applicable to the general news pages, which can extract the news paragraphs automatically, accurately and constantly. Our experimental results of several news sites show that our approach works well with a high accuracy rate over a long period of time.

As future work, we will modify our algorithm to improve the accuracy rate even further. We will use our approach to extract the news article contents from different kinds of the news sites and construct a global Web news system for discovering the useful information and knowledge from news sites.

References

1. American newspapers and the internet: Threat or opportunity? Technical report, The Bivings Group, July 2007.
2. AllInOneNews. http://www.allinonenews.com.
3. ChaSen. http://chasen-legacy.sourceforge.jp.
4. J. Chen and S.-C. Lui. Perception-oriented online news extraction. In *The Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 363–366, 2008.
5. D. de Castro Reis, P. B. Golgher, A. S. da Silva, and A. H. F. Laender. Automatic Web news extraction using tree edit distance. In *The Proceedings of the 13th International Conference on World Wide Web*, pages 502–511, 2004.
6. Y. Dong, Q. Li, Z. Yan, and Y. Ding. A generic Web news extraction approach. In *The Proceedings of the 2008 IEEE International Conference on Information and Automation*, pages 179–183, 2008.
7. F. Fukumoto and Y. Suzuki. Detecting shifts in news stories for paragraph extraction. In *The Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7, 2002.
8. Full-Text RSS. http://echodittolabs.org/fulltextrss.
9. Google News. http://news.google.com.
10. Y. Li, X. Meng, Q. Li, and L. Wang. Hybrid method for automated news content extraction from the Web. In *The Proceedings of the 7th International Conference on Web Information Systems Engineering*, pages 327–338, 2006.

11. B. Liu, P. V. Hai, T. Noro, and T. Tokuda. Towards automatic construction of news directory systems. In *The Proceedings of the 17th European-Japanese Conference on Information Modeling and Knowledge Bases*, pages 211–220, 2007.
12. B. Liu, H. Han, T. Noro, and T. Tokuda. Personal news RSS feeds generation using existing news feeds. In *The Proceedings of the 9th International Conference on Web Engineering*, pages 419–433, 2009.
13. Y. Lu, W. Meng, W. Zhang, K.-L. Liu, and C. Yu. Automatic extraction of publication time from news search results. In *The Proceedings of the 2nd International Workshop on Challenges in Web Information Retrieval and Integration*, page 50, 2006.
14. T. Noro, B. Liu, Y. Nakagawa, H. Han, and T. Tokuda. A news index system for global comparisons of many major topics on the earth. In *The Proceeding of the 18th European-Japanese Conference on Information Modeling and Knowledge Bases*, pages 197–213, 2008.
15. J. Parapar and A. Barreiro. An effective and efficient Web news extraction technique for an operational newsIR system. In *The Proceeding of XIII Conferencia de la Asociacion Espanola para la Inteligencia Artificial CAEPIA*, volume II, pages 319–328, 2007.
16. J. Prasad and A. Paepcke. CoreEx: Content extraction from online news articles. In *The Proceeding of the 17th ACM conference on Information and Knowledge Mining*, pages 1391–1392, 2008.
17. Readability. http://lab.arc90.com/experiments/readability/.
18. H. Shinnou and M. Sasaki. Automatic extraction of target parts from a Web page. In *IPSJ SIG Notes*, volume 2004-NL-162, pages 33–40, 2004. In Japanese.
19. Y. Shinyama. *Webstemmer*. http://www.unixuser.org/ẽuske/python/webstemmer/.
20. TidyRead. http://www.tidyread.com/.
21. J. Wang, X. He, C. Wang, J. Pei, J. Bu, C. Chen, Z. Guan, and G. Lu. News article extraction with template-independent wrapper. In *The Proceedings of the 18th International Conference on World Wide Web*, pages 1085–1086, 2009.
22. H. Zhao, W. Meng, and C. Yu. Automatic extraction of dynamic record sections from search engine result pages. In *The Proceedings of the 32nd International Conference on Very Large Data Bases*, pages 989–1000, 2006.
23. S. Zheng, R. Song, and J.-R. Wen. Template-independent news extraction based on visual consistency. In *The Proceedings of the 22th AAAI Conference on Artificial Intelligence*, pages 1507–1513, 2007.
24. C.-N. Ziegler and M. Skubacz. Content extraction from news pages using particle swarm optimization on linguistic and structural features. In *The Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 242–249, 2007.