An Effective and Efficient Web News Extraction Technique for an Operational NewsIR System

Javier Parapar and Álvaro Barreiro

IRLab, Department of Computer Science , University of A Coruña,
Campus de Elviña s/n, 15071, A Coruña, Spain
{javierparapar,barreiro}@udc.es
http://www.dc.fi.udc.es/irlab

Abstract. Web information extraction, in particular web news extraction is an open research problem and it is a key point in NewsIR systems. Current techniques fail in the quality of the results, the high computational cost or the necessity of human intervention, all of them critical issues in a real system. We present an automated approach to news recognition and extraction based on a set of heuristics about the articles structure, that is currently applied in an operational system.We also built a data set to evaluate web news extraction methods. Our results in this collection of international news, composed of 4869 web pages from 15 different on-line sources, achieved a 97% of precision and a 94% of recall for the news recognition and extraction task.

1 Introduction

The huge amount of news information available on-line requires the use of Information Retrieval (IR) techniques to avoid overwhelming the users. The main objectives of these IR methods are: reduce the time spend in reading the articles, avert the redundancy and provide topic search capability. In this context article recognition and extraction have become in key points.

The lack of news publication standards, the heterogeneous content of the web sites and the big amount of different publication formats make web news extraction an open problem. This subject was previously addressed adapting web data extraction techniques. Results were not satisfactory due to the low quality of the extraction process outcomes, the necessity of human intervention and the high computational complexity of these methods.

News recognition and news extraction approaches deal with two problems: the identification of news articles pages within a collection of heterogeneous web documents where there are many not desired content (e.g. section pages, headlines, advertisements...), and, given a document identified like a *news page*, the extraction of the fields of the article, i.e., the title, the body and the image, if present.

Existing approaches [6] to the problem consider both phases separately, which implies the need of multiple processing of the same document. The main problem of these methods is their complexity. The high computational cost arises from the adaptation of general data extraction techniques [2] and the use of clustering methods and algorithms based on the tree edit distance.

For a real system the computational cost is critical because thousands of sources are continuously updated. Poor computational behaviour would compromise the performance and scalability of the system.

To overcome these difficulties we propose a method for news recognition and extraction from a set of web documents, based on domain specific heuristics, that takes advantage of the common characteristics of the articles structure, resulting in an efficient and effective algorithm that solves the news data extraction problem.

Section 2 presents a short review of previous work in the area of news and data extraction. Section 3 introduces the system where our approach is currently applied. Section 4 explains the main features of our technique for article extraction and recognition. Section 5 shows the designed evaluation routine and the results. We propose future lines of work and conclude in Section 6.

2 Previous Work

As already explained a critical point in any web news information retrieval system is the method used to extract the articles. There is not too much research published in this field; only a few of works show approaches derived from the adaptation of data extraction general techniques to this task. Traditionally there are two opposite approaches to the recognition and extraction problem:

1. **Custom static templates**, mainly used by commercial solutions. In this approach for every source to be indexed by the system a extraction template has to be defined. The template indicates in which URLs of the publisher site are located the news articles. At web page level, the offsets or tags of the different article fields are defined with more or less flexibility. So in the extraction phase the pages of every publisher are individually processed filtering the documents that match with the URL criteria, for these pages the article parts are selected following the instructions of the template.

   The advantage of this kind of methods is the computational cost that allows to implement these approaches as efficient routines. In the cons part a lot of human intervention is needed. For every new source to be added to the system the administrator must analyse the internal HTML structure of the documents and define a custom template. If any of the publishers changes something in the publication format or the document structure, the template must be redefined. Thus the system maintenance becomes in a critical task.
2. **Automated extraction**. The most of the published works belong to this approach. The aim of these techniques is to avoid the human intervention and enable dynamically source adding to the systems.

As common characteristics all the algorithms take as input a set of web documents without any human tagging or specification; the process automatically identifies which documents are news and extracts from them the article fields as title, body, date, image etc. Although there is not too much research work there are mainly two ways to affront the automated solution:

- *Adaptation of data extraction traditional techniques* based on different clustering techniques as for example tree edit distance [6, 2] or use of equivalence classes [1]. The concept over these approaches lie, is that news with common structures will match in the same cluster or class, so after the clustering phase a extraction template could be generated for each cluster.This implies multiple reprocessing of the documents with prohibitive computational cost. For example, in the case of the tree edit distance method, the complexity is $O(n^4)$ where n is the size of the document trees. Thus this family of techniques is not applicable in real systems, it is only useful in applications where the number of documents managed is reduced and the frequency of content update is low.
- *Domain specific approaches*: Other approaches try to combine the previous knowledge in the area of data extraction taking in account the singular characteristics of the news domain. Some works try to exploit the structure of the articles by semantic partitioning [7] the HTML blocks present in the document, following by a taxonomy of the concepts and the labelling of the groups. This approach is not still computational efficient and the results of precision and recall claimed by the authors can be improved. Other recent work [8] tries to use the *tables* present in the documents after assume that the news are present in the larger cell, following by a content validation based on taking the first lines of the extracted text to make a keyword analysis. These keywords are used to check if the whole extracted text is coherent with the semantic of the first sentences. The authors claim that this validation works because these first lines usually are the title or the first paragraphs of the article. This approach has as main problem the basis assumption. Actually this assumption is false in most of the cases the news articles are not contained by tables. Also the evaluation methodology used in this work is very poor.

So in this context we present an *automated extraction* approach based on the exploitation, with a set of combined heuristics, of the *domain specific* characteristics. Our method is a tradeoff between computational efficiency and result effectiveness. We developed a linear algorithm on the document size that outperforms the quality of previous automated approaches.

3 The System

In the NewsIR systems field, although there are not too much published work, some commercial solutions are currently available. The idea of our development was to improve the features that these solutions offer to the user and add new ones that, as users, we missed in the existing systems. Of course this system, NowOnWeb[1], is also used as a research platform by the group to test our results on the information retrieval and data extraction and management areas.

3.1 Existing Solutions

News search has become in one of the most used internet services, for this reason most of the internet enterprises have developed their own news searchers. It is the case of services as Google News, Yahoo News or MSN Newsbot that offer to the user the possibility of searching in a vast news collection because of its computational capacity. But, at the time of writing this paper, the products are closed to the user and not allow personalisation or useful characteristics as summarisation or redundancy grouping.

In the commercial solutions the information is showed in a not natural way represented only by titles, this forces to the user to navigate to each publisher to get the correct meaning of the articles. We purpose a more comfortable navigation across the search result. The results are ranked based on its relevance to the user needs of information, redundant documents are grouped and a summary of each group representative is showed to the user.

The academic research is more centred in news clustering summarisation and events tracking than in article search and summarisation. The works of the Columbia University [3] are centred in *NewsBlaster* a system that provides multidocumental summarisation of related news about specific events. Michigan University also developed *NewsInEssence* [5] that offers news clustering search as main capability.

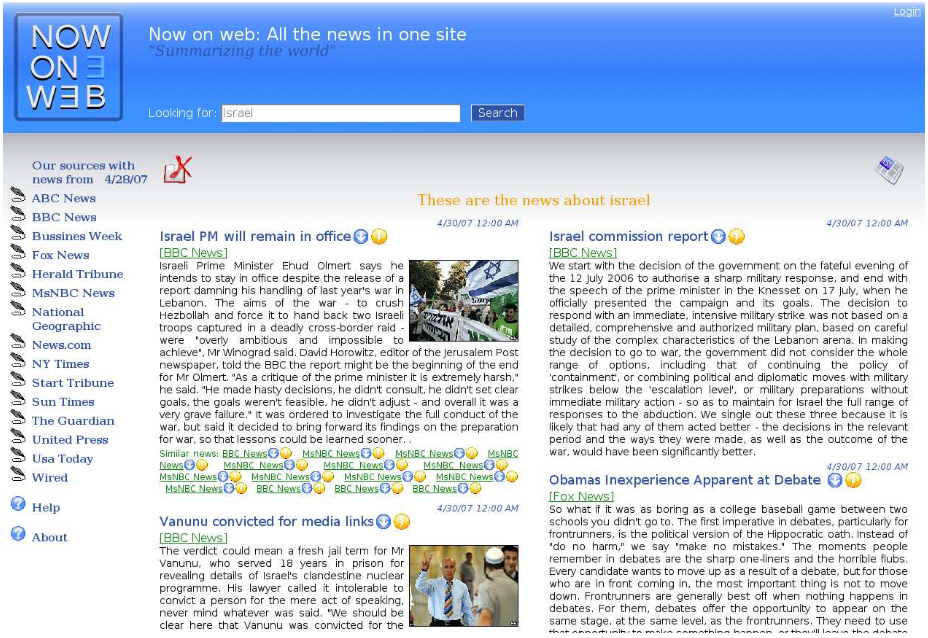
3.2 NowOnWeb

NowOnWeb was designed as a Model-View-Controller web-application following a component-based architecture. The main system components are: a crawler and an indexer to maintain an incremental index with a temporal window, a news recognition and extraction module that allows dynamic source adding, a news grouping component that uses a redundancy detection approach, and an article summariser based on the extraction of relevant sentences.

Our application offers the user: news searching among all the indexed publishers, query suggestion, redundancy detection and filtering, query biased summary generation, multiple format outputs like PDF or syndication services. All these characteristics aim to facilitate the use of the system, for this reason the results are showed in a friendly and natural way. In this sense technologies like AJAX were applied in order to improve the user experience and the system possibilities.

In Figure 1 a segment of the results to an user query (*"Israel"*) is showed. The articles relevant to the query are ranked and grouped to filter the

[1] NowOnWeb: http://nowonweb.dc.fi.udc.es



Login

Now on web: All the news in one site

"Summarizing the world"

Looking for: Israel Search

Our sources with news from 4/28/07

- ABC News
- BBC News
- Bussines Week
- Fox News
- Herald Tribune
- MsNBC News
- National Geographic
- News.com
- NY Times
- Start Tribune
- Sun Times
- The Guardian
- United Press
- Usa Today
- Wired

Help

About

These are the news about israel

4/30/07 12:00 AM

Israel PM will remain in office

[BBC News]

Israeli Prime Minister Ehud Olmert says he intends to stay in office despite the release of a report damning his handling of last year's war in Lebanon. The aims of the war - to crush Hezbollah and force it to hand back two Israeli troops captured in a deadly cross-border raid - were "overly ambitious and impossible to achieve", Mr Winograd said. David Horowitz, editor of the Jerusalem Post newspaper, told the BBC the report might be the beginning of the end for Mr Olmert. "As a critique of the prime minister it is extremely harsh," he said. "He made hasty decisions, he didn't consult, he didn't set clear goals, the goals weren't feasible, he didn't adjust - and overall it was a very grave failure." It was ordered to investigate the full conduct of the war, but said it decided to bring forward its findings on the preparation for war, so that lessons could be learned sooner. .

Similar news: BBC News MsNBC News MsNBC News MsNBC News MsNBC News MsNBC News MsNBC News MsNBC News MsNBC News MsNBC News MsNBC News MsNBC News BBC News BBC News BBC News

4/30/07 12:00 AM

Vanunu convicted for media links

[BBC News]

The verdict could mean a fresh jail term for Mr Vanunu, who served 18 years in prison for revealing details of Israel's clandestine nuclear programme. His lawyer called it intolerable to convict a person for the mere act of speaking, never mind whatever was said. "We should be clear here that Vanunu was convicted for the

4/30/07 12:00 AM

Israel commission report

[BBC News]

We start with the decision of the government on the fateful evening of the 12 July 2006 to authorise a sharp military response, and end with the speech of the prime minister in the Knesset on 17 July, when he officially presented the campaign and its goals. The decision to respond with an immediate, intensive military strike was not based on a detailed, comprehensive and authorized military plan, based on careful study of the complex characteristics of the Lebanon arena. In making the decision to go to war, the government did not consider the whole range of options, including that of continuing the policy of 'containment', or combining political and diplomatic moves with military strikes below the 'escalation level', or military preparations without immediate military action - so as to maintain for Israel the full range of responses to the abduction. We single out these three because it is likely that had any of them acted better - the decisions in the relevant period and the ways they were made, as well as the outcome of the war, would have been significantly better.

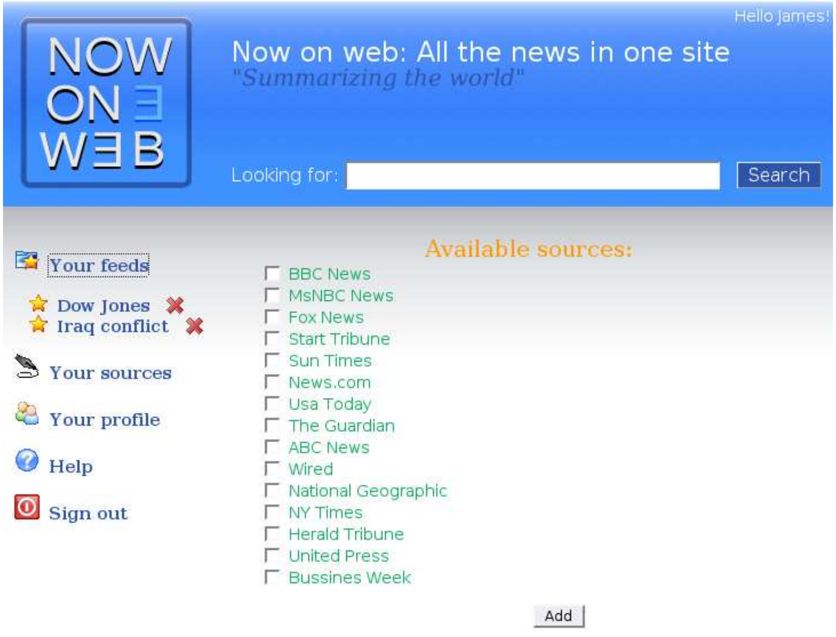
4/30/07 12:00 AM

Obamas Inexperience Apparent at Debate

[Fox News]

So what if it was as boring as a college baseball game between two schools you didn't go to. The first imperative in debates, particularly for frontrunners, is the political version of the Hippocratic oath. Instead of "do no harm," we say "make no mistakes." The moments people remember in debates are the sharp one-liners and the horrible flubs. Every candidate wants to move up as a result of a debate, but for those who are in front coming in, the most important thing is not to move down. Frontrunners are generally best off when nothing happens in debates. For them, debates offer the opportunity to appear on the same stage, at the same level, as the frontrunners. They need to use

Fig. 1. A snapshot of a the search results



Hello James!

Now on web: All the news in one site

"Summarizing the world"

Looking for: Search

Your feeds

- Dow Jones
- Iraq conflict

Your sources

Your profile

Help

Sign out

Available sources:

- □ BBC News
- □ MsNBC News
- □ Fox News
- □ Start Tribune
- □ Sun Times
- □ News.com
- □ Usa Today
- □ The Guardian
- □ ABC News
- □ Wired
- □ National Geographic
- □ NY Times
- □ Herald Tribune
- □ United Press
- □ Bussines Week

Add

Fig. 2. The user can choose the sources that contribute to the results

redundancy. The order of relevance is from left to the right and from top to bottom. For every redundancy group the most relevant document is selected to be shown to the user. For each article the system shows: the title, with a link to the original source, the summary, the image that illustrates the news, a link to the cached document as it was at crawling time and the list of the redundant documents with a link to the source, to the individual summary and to the cache version for every article in the group. Plus this a RSS feed to the query is available to subscription by the user, and a PDF report can be downloaded.

We have to notice that the summaries are created in retrieval time. In this sense the summaries are generated with a technique based on the extraction sentences relevant to the user query. So the generated summaries will reflect the article spirit centred in the needs of information of the user.

Figure 2 shows some of the personalisation options. An user can select which sources will contribute to his results. In this way we allow to the user select his information requirements filtering the sources that are interesting for him. The user also can save his frequent queries to consult them easily.

Summarising, these methods combined with an appropriate architecture and system design resulted in a NewsIR system that satisfies the user needs of information, allowing them to be up-to-date without time waste. We got an original solution different from the existing ones in the academic and commercial fields.

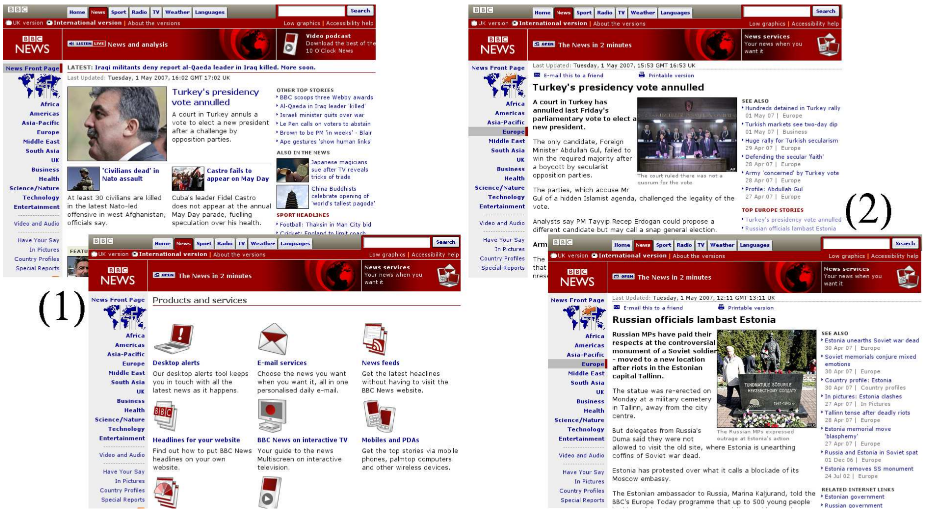
4 The News Recognition and Extraction Method

News recognition and extraction are considered together by our heuristics. Although the working collection is only composed of web pages from news sites, there are many pages that are not *news pages* such as section pages, or pages containing headlines, galleries, advertisements, etc.. See Figure 3 for examples and counterexamples of news pages. News data extraction addresses the problem of obtaining the different parts of a news article, that are the title, the news body, and the image.

We propose a set of heuristics to identify and extract the article, or reject a not-news web page:

1. *News are composed of paragraphs that are next one each other, although it could appear some not-desired content between them.*
2. *Paragraphs have a minimum size. News also have a minimum size.*
3. *Paragraphs are mostly text. Only styling markup and hyperlinks are allowed in paragraphs. Markup abounds in not desired content.*
4. *A low number of hyperlinks are allowed in paragraphs. A high number indicates not desired content, probably section pages.*

We also developed a set of heuristics to identify the title of the articles and the image, if it exists. Basically the title is on the top of the news body and has a special typing style, usually it also substantially matches with the content of the HTML *title* tag. Among all the images present in a web page we have to choose which one is relevant with the article;



Video podcast
Download the best of the 10 O'Clock News

LATEST: Iraqi militants deny report al-Qaeda leader in Iraq killed. More soon.

Last Updated: Tuesday, 1 May 2007, 16:02 GMT 17:02 UK

Turkey's presidency vote annulled

A court in Turkey annuls a vote to elect a new president after a challenge by opposition parties.

'Civilians dead' in Nato assault

At least 30 civilians are killed in the latest Nato-led offensive in west Afghanistan, officials say.

Castro fails to appear on May Day

Cuba's leader Fidel Castro does not appear at the annual May Day parade, fuelling speculation over his health.

OTHER TOP STORIES
- BBC scoops three Webby awards
- Al-Qaeda in Iraq leader 'killed'
- Israeli minister quits over war
- Le Pen calls on voters to abstain
- Brown to be PM 'in weeks' - Blair
- Ape gestures 'show human links'

ALSO IN THE NEWS

Japanese magicians sue after TV reveals tricks of trade

China Buddhists celebrate opening of 'world's tallest pagoda'

SPORT HEADLINES
- Football: Thaksin in Man City bid
- Cricket: England to limit coach

FEATU

News services
Your news when you want it

Products and services

Desktop alerts
Our desktop alerts tool keeps you in touch with all the latest news as it happens.

E-mail services
Choose the news you want when you want it, all in one personalised daily e-mail.

News feeds
Get the latest headlines without having to visit the BBC News website.

Headlines for your website
Find out how to put BBC News headlines on your own website.

BBC News on interactive TV
Your guide to the news Multiscreen on interactive television.

Mobiles and PDAs
Get the top stories via mobile phones, palmtop computers and other wireless devices.

Last Updated: Tuesday, 1 May 2007, 15:53 GMT 16:53 UK

Turkey's presidency vote annulled

A court in Turkey has annulled last Friday's parliamentary vote to elect a new president.

The only candidate, Foreign Minister Abdullah Gul, failed to win the required majority after a boycott by secularist opposition parties.

The court ruled there was not a quorum for the vote

The parties, which accuse Mr Gul of a hidden Islamist agenda, challenged the legality of the vote.

Analysts say PM Tayyip Recep Erdogan could propose a different candidate but may call a snap general election.

Arm

The that pres

SEE ALSO
- Hundreds detained in Turkey rally
  01 May 07 | Europe
- Turkish markets see two-day dip
  01 May 07 | Business
- Huge rally for Turkish secularism
  29 Apr 07 | Europe
- Defending the secular 'faith'
  28 Apr 07 | Europe
- Army 'concerned' by Turkey vote
  28 Apr 07 | Europe
- Profile: Abdullah Gul
  27 Apr 07 | Europe

TOP EUROPE STORIES
- Turkey's presidency vote annulled
- Russian officials lambast Estonia

Last Updated: Tuesday, 1 May 2007, 12:11 GMT 13:11 UK

Russian officials lambast Estonia

Russian MPs have paid their respects at the controversial monument of a Soviet soldier - moved to a new location after riots in the Estonian capital Tallinn.

The statue was re-erected on Monday at a military cemetery in Tallinn, away from the city centre.

But delegates from Russia's Duma said they were not allowed to visit the old site, where Estonia is unearthing coffins of Soviet war dead.

The Russian MPs expressed outrage at Estonia's action

Estonia has protested over what it calls a blockade of its Moscow embassy.

The Estonian ambassador to Russia, Marina Kaljurand, told the BBC's Europe Today programme that up to 500 young people

SEE ALSO
- Estonia unearths Soviet war dead
  30 Apr 07 | Europe
- Soviet memorials conjure mixed emotions
  30 Apr 07 | Europe
- Country profile: Estonia
  30 Apr 07 | Country profiles
- In pictures: Estonia clashes
  27 Apr 07 | In Pictures
- Tallinn tense after deadly riots
  28 Apr 07 | Europe
- Estonia memorial move 'blasphemy'
  27 Apr 07 | Europe
- Russia and Estonia in Soviet spat
  01 Dec 06 | Europe
- Estonia removes SS monument
  24 Jul 02 | Europe

RELATED INTERNET LINKS
- Estonian government
- Russian government

Fig. 3. Examples of not-news pages (1) and news pages (2) present in the news web sites indexed

for this task we consider the position next or inside the news body, the content of the HTML *alt* tag, the size of the picture, the image format and the *URL*.

The implementation of these heuristics results in a linear complexity algorithm on the web page length. The algorithm follows the content of the HTML documents looking for paragraphs that match with the criteria and joins them to build the news body. It also requires to set-up the values of some parameters (paragraph minimum size, news body minimum size, inter-paragraph maximum distance and hyperlink density), but in fact they were easy to set with minimum experimentation with daily news because they are stable among the articles published. These parameters where previously tunned over a hold-out data set.

5 Evaluation of the Method

In the web news extraction area there is no standard collection to compare the different methods. Although there are several news collections in the field of the TREC [4] initiative, they are not valid for our purpose because are composed only of plain text with the articles already extracted and tagged. So we carefully built up our own data set for evaluation . We also designed an evaluation routine capable to determine the goodness of our method.

5.1 The Data Set

We got a data set of 4869 web pages crawled from 15 different international top level on-line news sources[2]. These pages were crawled and indexed from the web with depth 3 and without any filtering or manual discarding of anyone. In this way we built a collection that is very representative of the reality of the on-line news area.

5.2 The Evaluation Routine and Results

The use of publishing systems implies that URLs share a common structure. For example, the following is a regular expression that matches the most of news pages for BBC News: *http://news.bbc.co.uk/.*/[0-9]+.stm*. This information is not useful as an heuristic for news recognition because publishers could modify the URLs at any moment, but the evaluation process can be benefited from knowing these patterns.

The evaluation routine has two phases. In a first step we run our algorithm that implements our heuristics against the data set. Result of this step we record the set of news articles obtained. We compare this extracted articles set with the the set of news extracted by URL pattern criteria. If a page is in both sets is a true positive, if it is not in any of them is a true negative.

After this stage the false positives (the news extracted by the algorithm with an URL that does not match with the pattern criteria) and false negatives (the that have not been extracted using our heuristics but which URL match with the pattern criteria) are manually inspected to determine whether they are news articles or not.

For the manual revision we designed an application where false negatives and false positives web pages are randomly showed to the human assessor. The assessor has no information about the previous judgements and only must decide whether there is a news article in the page or not.

From the results of the manual revision, we could determine that most of the news articles not extracted by our algorithm it was because they were very short press notes. In the case of extracted not-articles it was because they were titled long abstracts or editorials.

As result of both phases we assessed that the collection is composed of 2658 web pages with articles and 2211 non articles pages. The results of the evaluation of our heuristic for news extraction are the following:

- From the 2658 articles in the data-set the method was able to extract 2506 whereas the other 152 were rejected, mainly because there were very short news notes.
- From the 2211 non-articles pages only 70 where accepted and extracted as articles, the most of them containing big news summaries and digests.

[2] ABC News, BBC News, Business Week, Fox News, Herald Tribune, MsNBC News, CNET News.com, National Geographic, New York Times, Start Tribune, Sun Times, The Guardian, United Press, Usa Today, Wired.

In this context to asses the results of our method we use the concepts of precision and recall. Precision has here the meaning of the number of articles correctly recognised divided by the sum of the correctly recognised plus the non-articles recognised and recall means the number of articles correctly recognised divided by the sum of the correctly recognised plus the articles not recognised. Taking these definitions, finally our method values of precision and recall were 97,28% and 94,28% respectively.
These results outperforms the ones of previous approaches. As we explained previously the absence of standard test collections or a defined methodology make difficult the comparison. Anyway we can compare precision and recall computed in different datasets, but that are similar because are composed of web pages from news sites. In this terms for example the work Vadrevu et al. [7] uses a data set composed by American web news sites [3]. In this collection of 3216 pages and 12 different publishers their method obtained 87% precision and 85% recall.

6 Conclusions and Future Work

In this paper we have introduced a method for web news recognition and extraction that takes advantage of the domain specific characteristics. Our algorithm is based on a set of heuristics and its complexity is linear on the document size.
We also designed an original evaluation strategy and built up a data set that could be used by others. The evaluation assessed that our approach obtains very high values of precision and recall. This method is currently working in an operative news system
This method is used because it is easy to tune the parameters implied in the heuristics. A challenge that we are considering is adapt our algorithm to other fields and tasks where the content to recognise would be much more variable. Studies about the stability of the parameters and the use of statistical learning approaches to estimate the best parameter values would be necessary in these domains.

7 Acknowledgements

This work was cofunded by the “Secretaría de Estado de Universidades e Investigación” and FEDER funds (MEC TIN2005-08521-C02-02) and “Xunta de Galicia”(PGIDIT06 PXIC10501PN).

References

1. A. Arasu, H. Garcia-Molina, and S. University. Extracting structured data from web pages. In *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 337–348, New York, NY, USA, 2003. ACM Press.

[3] ABC News, BBC News, Fox News, MsNBC News, CBC News, CBS News, CNN News, New York Times, Reuters, Time.com, Times Online, Usa Today

2. V. Crescenzi and G. Mecca. Automatic information extraction from large websites. *J. ACM*, 51(5):731–779, 2004.
3. K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *Proceedings of the Human Language Technology Conference*, 2002.
4. National Institute of Standards and Technology. TREC: Text REtrieval Conference. *http://trec.nist.gov*, 2007.
5. D. Radev, J. Otterbacher, A. Winkel, and S. Blair-Goldensohn. NewsInEssence: Summarizing Online News Topics. *Commun. ACM*, 48(10):95–98, 2005.
6. D. C. Reis, P. B. Golgher, A. S. Silva, and A. F. Laender. Automatic web news extraction using tree edit distance. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 502–511, New York, NY, USA, 2004. ACM Press.
7. S. Vadrevu, S. Nagarajan, F. Gelgi, and H. Davulcu. Automated metadata and instance extraction from news web sites. In *WI '05: Proceedings of the The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pages 38–41, Washington, DC, USA, 2005. IEEE Computer Society.
8. D. Zhang and S. J. Simoff. Informing the curious negotiator: Automatic news extraction from the internet. In G. J. Williams and S. J. Simoff, editors, *Selected Papers from AusDM*, volume 3755 of *Lecture Notes in Computer Science*, pages 176–191. Springer, 2006.