

Perception based Web News Extraction

Jinlin Chen
Computer Science Department
Queens College
City University of New York
Flushing, NY 11367
jchen@cs.qc.edu

Keli Xiao
Computer Science Department
Queens College
City University of New York
Flushing, NY 11367
keli.xiao@gmail.com

ABSTRACT

A novel online news extraction approach based on human perception is presented in this paper. The approach simulates how a human perceives and identifies online news information. It first detects news content areas based on content function, space continuity, and formatting continuity of news information. It further identifies detailed news content based on the position, format, and semantic of detected news areas. Experiment results show that our approach achieves much better performance (in average more than 99% in terms of F1 Value) compared to previous approaches such as Tree Edit Distance and Visual Wrapper based approaches. Furthermore, our approach does not assume the existence of Web templates in the tested Web pages as required by Tree Edit Distance based approach, nor does it need training set as required in Visual Wrapper based approach. The success of our approach demonstrates the strength of the perception based Web information extraction methodology and represents a promising approach for automatic information extraction from sources with presentation design for humans.

Categories and Subject Descriptors

H.3.m [Miscellaneous]: Information extraction, Web

General Terms

Algorithms, Experimentation

Keywords

Information extraction, online news, Web

1. INTRODUCTION

Information extraction has become an important technology to help users locate desired information on the Web. Designing a generalized method for extracting Web information is complicated due to the heterogeneity of Web information. Because of this, domain specific characteristics are often considered for effective Web information extracting. One such domain is on-line news websites. With thousands of new websites to provide daily news in today's Web, it is critical to provide a tool that can automatically extract online news information for users.

Most of previous approaches use manually or automatically

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 2006 ACM. \$5.00.

constructed wrappers to extract news information. Such approaches assume that news information is wrapped by recurring physical or virtual patterns across news pages. Their major task is to find or learn appropriate wrappers. One such approach is Tree Edit Distance (refer as TED below) [7] which generates wrappers based on the consistency of HTML DOM trees. Another approach is visual wrapper (refer as VW below) [9] based, which learns wrappers based on recurring visual patterns.

Several problems exist in previous approaches for online news extraction. First, they have special prerequisites which may limit their generalized usage. For example, TED is template-dependent and requires that multiple pages with the same templates exist in the news corpse to be extracted. VW requires a training stage to derive wrappers. The training data need be labeled manually. Extraction results may not be satisfactory when training set is too small. Second, even with these prerequisites satisfied, the extraction results may still be unstable and domain/site dependent. The major reason is that they have inappropriate assumptions which may not always hold due to the heterogeneous and dynamic characteristics of the online news information. For example, TED assumes that templates exist and can be extracted based on DOM trees, thus the generated wrapper can only work properly for pages that share a specific template that recurs in DOM trees of Web pages. Violation to this (which is quite normal practically) will lead to the invalidation of a wrapper. VW assumes that news content has the following visual features: (1) Its area is larger than surrounding objects; (2) A bold-faced news headline exists at its top; (3) It is occupied by contiguous text paragraphs, sometimes mixed with one or two pictures. (4) Its center is close to that of the whole page. Even though these assumptions may hold in many cases, they are not always true due to the diversity of web authoring technique, news content property, purpose of the news content, as well as the various types of noise information interwoven with news contents. Many exceptions exist and result in the ineffectiveness of the approach.

Despite the heterogeneous nature of Web pages, humans are very effective at identifying news information, even when they do not understand the language or content of the news. The major reason is that news pages are designed for humans. Even though the information, format, and layout may change in various ways, the presentation design as a whole should be easily recognized by human readers. As a result human users generally have no problem in recognizing news contents based on their perception.

The motivation of our research is to identify how humans perceive and recognize news content, and simulate such mechanism to create an effective news extraction algorithm that is stable across any presentation designs and news domains.

Given a news page, a human reader first scans the page to identify the areas that contain the news content. News areas generally have some special properties that distinguish them from other areas, 1) Functional property: the function of a news content area is mainly informational; 2) Space continuity: information that belongs to a news content area should be located continuously in space (vertically and/or horizontally), and generally they may only be separated by content areas that are not pure text based information areas; 3) Formatting continuity: major formats of different news content areas should be similar. Even though each of these properties may not always be consistent, the combinations of them generally provide sufficient criteria to identify news content areas. In this paper we aim at developing an effective algorithm to detect news content areas based on these properties.

After detecting the news areas, the reader will further read content within the area in a second scan. In this scan the reader will identify precisely which information is news in the major news content area. Note that most of the properties that the reader uses to decide the content areas are still valid to decide whether a component is really news, besides, semantic is also a property that the reader uses at this stage. Errors may happen if a user does not understand the semantic of a component in the news area.

In this paper we present an effective approach for online news extraction which simulates how humans perceive and recognize news information based on the analysis above. Experiment results show that our approach performs much better than previous approaches such as TED and VW. Besides, our approach does not assume the existing of Web templates in the tested websites as required by TED, nor does it need training set as required by VW.

The success of our approach demonstrates the strength of the perception based Web information extraction methodology and represents a promising approach for automatic information extraction from sources with presentation design for humans.

The remainder of the paper is structured as follows: we first briefly review related works in section 2. We then present our approach and evaluate its performance in Section 3 and 4. Finally, we conclude the paper in Section 5.

2. RELATED WORKS

News extraction is a special topic in the area of Information Extraction (IE). Today the major efforts on Web IE focus on automatic training approach which has two main categories, wrapper-based approach [2][6] where rules are automatically discovered using predefined templates, and the statistical generative model where statistical models or parsers are constructed and then decoded to find relevant content. Hidden Markov Models [3][8][10] (HMMs) have been among the most accurate methods for automatic training approach. In [5] a brief surveys on IE is given.

Most approaches on news extraction focus on wrapper based approaches. Reis et al. [7] presented a tree edit distance based approach to generate extraction patterns from clustered pages. Pages in each cluster share a common template. One major problem with this approach is that it is template-dependent. Zheng et al [9] presented a template-independent news extraction approach based on visual consistency. The approach first represents a page as a visual block tree, and then derives a composite visual feature set that is stable in the news domain by extracting a series of visual features. Finally, machine learning

approach is used to generate a vision-based wrapper, which is used to extract news information. One major problem of this approach is that it needs manually labeled training data set, and if the training set is not large enough, the extraction result may be inaccurate.

3. PERCEPTION BASED ONLINE NEWS EXTRACTION

One major property humans use to identify online news is content functions, i.e., the intended usage of the content. In our previous work [1][4], a Function based Object Model (FOM) has been presented to detect the function of Web content. In FOM each piece of content in a Web page is represented as an Object with certain functions. A Basic Object does not contain any other Objects while a Composite Object contains a set of Objects that perform some certain functions together. There are four major types of Object based on their major functions, 1) Information Object; 2) Navigation Object; 3) Interaction Object; and 4) Decoration Object. An Object with two or more major functions is a Mixed Object. A Mixed Object generally contains children with different function types. For news extraction purpose, we further define the following types of Objects based on FOM.

Definition 1 (Block Object and Inline Object). A Block Object is an Object displayed as independent content blocks separated from other blocks by vertical and horizontal spaces. Inline Objects are usually displayed one after another on a line within the Block Object that contains them.

Definition 2 (Leaf Block Information Object). A Leaf Block Object is a Block Object that does not contain any other Block Object, i.e., it can only contain Inline Objects. A Leaf Block Information Object (LBIO) is a Leaf Block Object whose major function is to provide information. A Leaf Block Information Object whose major media type is text is a Text Leaf Block Information Object (TLBIO). Otherwise it is a Media Leaf Block Information Object.

Definition 3 (Text/Media Information Object). An Information Object whose major content media type is text is called a Text Information Object. Otherwise it is a Media Information Object.

In this paper we focus on text news extraction, the approach can be easily extended to include news information in other media types. Next we present relevant axioms and corollaries as a starting point to simulate human perception for news extraction.

Axiom 1. News content in a news Web page is a set of TLBIOs.

This axiom states the fact that the major function of news content is to provide information. Therefore, Navigation, Interaction, and Decoration Objects are not taken as news content. Based on this axiom, we use LBIO as the basic unit for news extraction. This conforms to the way human perceives news information.

Axiom 2. A news TLBIO can only be contained in an Information or Mixed Object.

Corollary 1. If a Leaf Block Object is contained in another Object whose main function is navigation, interaction, or decoration, it is not a news Object.

Proof. Given a Navigation, Interaction, or Decoration Object x , if it contains a Leaf Block Object c , and c is a news Object, then based on Axiom 1, c is a TLBIO. Based on Axiom 2, c can only

be contained in an Information or Mixed Object. This contradicts with the fact that x is not an Information or Mixed Object. Therefore, c cannot be a news Object.

Axiom 3. News TLBIOs of a news page are presented in one or more rectangular areas. Vertically, these rectangular areas are separated by Media Information Objects and/or non-Information Objects.

Corollary 2. Given two horizontally overlapped news areas a and b , if a and b are vertically separated by a Text Information Object c , then c is also news content, and we can merge a , c , and b into a bigger news Object.

Proof. If c is not news content, then a and b are not separated by a Media Information Object or a non-Information Object. This contradicts with Axiom 3. Therefore, c is also news content.

Generally speaking, vertical gap in between two vertically consecutive TLBIOs inside a news area is smaller than that in between a news area and its vertically adjacent non-news area.

Axiom 4. Contents in different news areas have similar major formats which are the presentation format used by the majority of Objects inside news areas

This axiom states the consistency of major format among news Object, which conforms to human perception.

Based on the axioms and corollaries above, given a news web page, we can first detect all its TLBIOs, merge them to derive possible news areas, and then verify each TLBIO based on their position, format, and semantic relevance to the news areas to detect all the news TLBIOs. Fig. 1 presents our algorithm for news extraction. The algorithm accepts a news web page p as input and generates news content as follows:

1) FOM analysis. We first perform FOM analysis on p to generate a FOM page f_p based on our previous work [1][4]. f_p is a tree based hierarchical structure representing the logically containing relationship of Objects with their functions detected;

2) TLBIO Detection. This subroutine accepts f_p as input and generates the set of all the TLBIOs in p . Based on Axiom 1, 2, and Corollary 1, it recursively check the children of f_p that are Information or Mixed Object and contain other Block Objects to detect TLBIOs. We also check the children of any Composite Block Object whose area ratio to the page is sufficiently big ((practically we use 40% as threshold) regardless of its function type. The reason is that functional analysis is generally more accurate on smaller areas than larger areas. Checking the children of large areas can avoid missing an Information Object.

3) News areas detection. This subroutine accepts the detected TLBIO set as input and recursively merge vertically adjacent areas with small gaps or similar formats based on Axiom 3, 4, and Corollary 2. One major feature of our approach is that we use adaptive minimum gap value to merge adjacent areas until the total number of merged areas is smaller than area number threshold. At the end of each round of merge, if the total number of merged areas is larger than the threshold, we will increase the value of minimum gap and repeat merge. The subroutine then decides major news area based on text size, hyperlink property, position and other features, and finally derives news areas based on whether their formats are similar to that of the major formats.

Note that the detected news areas may be more than one if the news page contains news in multiple columns.

4) News Detection. This subroutine accepts the detected news areas and TLBIO Set as input and derives news information based on the position, format, and/or semantic of each TLBIO.

5) Header detection. This is a relatively simple task because of the special features of titles. A title is generally a TLBIO with less than 20 words and close to the news body. It generally has the largest font size in the neighboring news areas. Semantically a title is similar to that of the news content. Our experiments results show that these heuristic are extremely effective in title detection.

Algorithm 1. Perception based news extraction

Input: A News Web page p

Output: Text news content in p

Method: *newsExtraction(p)*{

$f_p = FOMAnalysis(p)$;

$TLBIOSet = TLBIODetection(fp)$;

$areas = newsAreasDetection(TLBIOSet)$;

$news = newsDetection(areas, TLBIOSet)$;

$header = headerDetection()$;

insert header to the beginning of news;

return news;

}

Subroutine TLBIODetection (fp){

$TLBIOSet = \emptyset$;

for each block level child c of fp {

if c contains other Block Object{

if c is Information or Mixed Object, TLBIODetection (c);

else if c' area ratio > ARThreshold, TLBIODetection (c);

}

else if c is a Text Information Object c → TLBIOSet;

}

return TLBIOSet;

}

Subroutine newsAreasDetection (IOSet){

Sort Objects in IOSet based on their top position;

$MinGap = 0$;

do{

Merge horizontally overlapped and vertically adjacent blocks whose vertical gap is less than MinGap;

Merge horizontally overlapped and vertically adjacent blocks with similar formats;

$MinGap += GapIncrement$;

while (|IOSet| > AreaNoThreshold);

$majorNewsIO = majorAreaDetection(IOSet)$;

$newsAreaSet = \emptyset$;

for each Object o in IOSet{

if o has similar format as majorNewsIO;

$newsAreaSet \leftarrow o$;

}

return newsAreaSet;

}

SubroutineneedsDetection(areas, TLBIOSet){

$newsInfo = \emptyset$;

for each TLBIO c in TLBIOSet{

If c is contained in news areas and c has similar format as major format or similar semantic as news areas,

$c \rightarrow newsInfo$;

```

    if c is partially contained in news areas and c has
    similar format and semantic as major format and
    semantic of news areas, c→newsInfo;
}
return newsInfo;
}

```

Figure 1. Algorithm for perception based news extraction

4. PERFORMANCE EVALUATION

A total of 745 pages from 19 news sites were tested for performance evaluation. To make our evaluation fair to other algorithms as well as comprehensive, we select the first 15 sites from the testing sites of [7], and added 4 other popular news sites (BBC, CNN, NYTimes, and AOL) for diversity.

Figure 2 compares the experiment results of Perception based approach and Tree Edit Distance based approach site by site on the first 15 news sites. The Y-axis *F1* value represents the harmonic mean of precision and recall. *F1* is defined as $2PR/(P+R)$, where *P* (Precision) is the number of correctly extracted news pages divided by the total number of extracted news pages, and *R* (recall) is the number of correct extracted pages divided by the total number of pages that contain news.

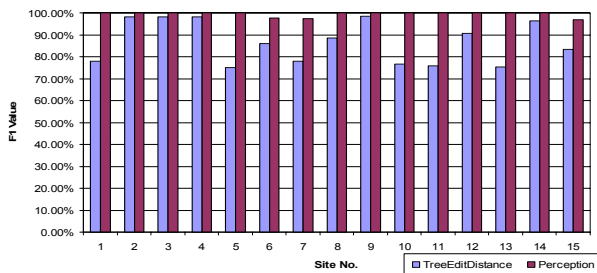


Figure 2. Experiment results different sites

Fig. 2 shows that Perception based approach performs much better than Tree Edit Distance based approach. The former outperforms the latter in every website tested. The average *F1* value of Perception based approach on the 15 websites tested is 99.46% (99.59% on the 19 sites tested), which is much higher than that of Tree Edit Distance based approach (86.45%).

The performance of VW depends on the size of training set which gives VW prior knowledge about testing website. Our approach, on the contrary, does not need any prior knowledge. Based on the testing results published by the authors of VW, initially, the first V-Wrapper learned from one site's training set can only achieve 50.93% accuracy in terms of *F1* Value, and the accuracy improves gradually when more training data is giving, the best performance (95%) comes when the training data from all site are applied to VW before apply it to test, which is still lower than our approach. Additionally, in practice we cannot expect training data from all the testing websites are given, which is too time consuming and also too constraint because in some applications the testing websites may even be unknown before hand.

The success of our approach is mainly because that we simulate the way humans perceive and identify news content, which is quite stable because news pages are designed for human to read. VW derives wrappers based on assumptions on visual information

such as position and size of news areas, which may be invalid due to the inconsistency of visual design and noise information.

5. CONCLUSIONS

In this paper we proposed a novel online news extraction approach based on human perception. The approach simulates how a human perceives and identifies online news information. Experiments show that our approach achieves extremely high accuracy (in average more than 99%), which is much better comparing to previous approaches such as TED (86.45%) and VW (at best 95% when training data from all the testing websites are available, but may be much lower with less training data). Furthermore, our approach does not assume the existence of Web templates in the tested Web pages as required in TED, nor does it need training set as required in VW.

We believe that the implication of this method is far beyond yet another efficient online news extraction algorithm. It demonstrates the strength of the perception based Web information extraction methodology and represents a promising approach for the applications that automatically extract information from sources with presentation design for human. In the future we expect to extend our approach for generalized Perception based Web information extraction.

6. REFERENCES

- [1] Chen, J., Zhou, B., Shi, J., Zhang, H., & Wu, Q. 2001. Function-based Object Model Towards Website Adaptation. Proc. of WWW-10. 587-596.
- [2] Freitag, D., & Kushmerick, N. 2000. Boosted wrapper induction. AAAI/IAAI 2000. 577-583.
- [3] Geng, J. and Yang, J. 2004. Automatic extraction and integration of bibliographic information on the Web. IDEAS '04. 193-04.
- [4] Gu, X., Chen, J., Ma, W., & Chen, G. 2002. Visual Based Content Understanding towards Web Adaptation. 2nd Intl. Conf. on Adaptive Hypermedia and Adaptive Web Based Systems. 164-173.
- [5] Laender, A. H. F.; Ribeiro-Neto, B. A.; da Silva, A. S.; and Teixeira, J. S. 2002. A brief survey of web data extraction tools. SIGMOD Record 31(2):84-93.
- [6] Muslea, I., Minton, S., Knoblock, C. 2001. Hierarchical Wrapper Induction for Semistructured Information Sources. Journal of Autonomous Agents and Multi-Agent Systems, 4(1/2), 93-114.
- [7] Reis, D. C., Golgher, P. B., Silva, A. S., and Laender, A. F. 2004. Automatic web news extraction using tree edit distance. In WWW2004, 502 - 511.
- [8] Skounakis, M., Craven, M., & Ray, S. 2003. Hierarchical hidden markov models for information extraction. Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence.
- [9] Zheng S., Song R., and Wen J. 2007. Template-independent news extraction based on visual consistency. AAAI-2007, 1507 -- 1513.
- [10] Zhong P. and Chen J. 2008. Web Information Extraction Using Web-Specific Features, Journal of Digital Information Management (to appear).