Informing the Curious Negotiator: Automatic News Extraction from the Internet

Debbie Zhang and Simeon J. Simoff

Faculty of Information Technology,
University of Technology, Sydney
Broadway PO Box 123, NSW 2007, Australia
{debbiez, simeon}@it.uts.edu.au

Abstract. In negotiation, information acquisition and validation play an important role in the decision making process. In this paper we briefly present the framework of a smart data mining system for providing contextual information from the Internet to a negotiation agent. We then present one of its components in more details - an effective automated technique for extracting relevant articles from news web sites, so that they can be used further by the mining agents. Most current techniques experience difficulties to cope with changes in websites structure and formats. The proposed extracting process is completely automatic and independent of web site formats. The technique is based on identifying regularities in both format and content of the news web sites. The algorithms are applicable to both single- and multi-document web sites. Since invalid URLs can cause errors in data extraction, we also present a method for the negotiation agent to estimate the validity of the extracted data based on the frequency of the relevant words in the news title. This paper also presents a new procedure for constructing news data sets of given topics. The extracted news data set is further utilised by the parties involved in negotiation. The information retrieved from the data set can support both human and automated negotiators.

1 Introduction

The *curious negotiator* [1] is a multiagent system of competitive agents supporting multi-attribute negotiation where the set of issues is not fixed [2]. The overall goal of its design is to exploit the interplay between contextual information [3] and the development of offers in negotiation conducted in an electronic environment. Current design is illustrated in Figure 1. Negotiation agents apply the negotiation strategies in the negotiation process [4]. With respect to the curious negotiator the term 'negotiation strategies' includes strategies for developing the set of issues in an offer as well as *identifying, requesting and evaluating contextual information* including determining what information to table as the negotiation proceeds [5]. A negotiation strategy should generally rely on information drawn from the context of the negotiation. The significance of information to the negotiation process was analysed formally in the seminal paper by Milgrom and Weber [6] in which the Linkage Principle, relating the revelation of contextual information to the price that a purchaser is prepared to pay, was introduced. "Good negotiators, therefore, undertake integrated processes of knowledge acquisition that combine sources of knowledge obtained at and away from the negotiation table. "They learn in order to plan and plan in order to learn" [7]. The grand vision for curious negotiator encapsulates this observation. The mediation agents (labelled as 'mediator' in Figure 1) assist negotiation agents in the negotiation process. The role of observer agents (labelled as 'observer' in Figure 1) is to observe and analyse what is happening on the 'negotiation table' and to look for opportunities particularly from failed negotiations.

Successful negotiation relies on an understanding of how to 'play' the negotiation mechanism [5] and on contextual information. From a process management point of view, negotiation processes are interesting in that they are knowledge-driven emergent processes that can be fully managed provided that, first, full authority to negotiate is delegated to the agent and, second, sufficient contextual information can be derived from the market data, from the sources, available on the Internet (news feeds, company white papers, specialised articles, research papers) and other sources by the data mining bots. The dashed lines in Figure 1 contour two scenarios: "SA" – a semi-automated scenario in which the human agent receives and processes contextual information and affects the strategies of the negotiation agent, and "A" in which contextual information is distilled and passed to the negotiation agent in a form of parameters that are taken in consideration by the negotiation strategies. The curious negotiator is designed to incorporate data mining and information discovery methods [8] that operate under time constraints, including methods from the area of topic detection and event tracking research [9]. The idea is encapsulated in the "smart data miner" in Figure 1. The architecture of this specialised data mining system, which operates in tandem with the human or/and negotiation agent, is shown in Figure 2. Initially the information is extracted from various sources including on-line news media, virtual communities, company and government web sites. Extracted information is converted to a structured representation and then both representations are stored in the mining base. They are used for further analysis by different data mining algorithms, including different text and network mining agents. The 'Source profile base', includes a collection of time-stamped data about the behaviour of the approached sources like response time, the number of answered requests, dates when a new layout appears, redirections of requests, types of errors, subscription price, change in subscription price, change of the level of service provided and other parameters. The 'Source evaluator' provides a number of estimates, e.g. 'hold-up', 'reliability', 'cost', 'trust' that evaluate the quality of the data sources from which the patterns have been extracted. These estimates are derived from the related data in the source profile base. This paper is limited to the techniques that cover the automatic extraction of relevant news articles. Regardless of whether we deal with scenario "SA" or "A", there are a number of challenges in real world negotiations that the smart data mining system needs to address, including (i) critical pieces of information being held in different repositories; (ii) non-standard formats; (iii) changes in formats at the same repository; (iv) possible duplicative, inconsistent and erroneous data. This paper addresses the first three challenges in the context of providing news to the negotiation table. The scope of the paper covers the universal news bot shown within the dashed rectangle in Figure 2. The techniques considered in this paper are applicable for both scenarios, however, the details are beyond the scope of the paper.

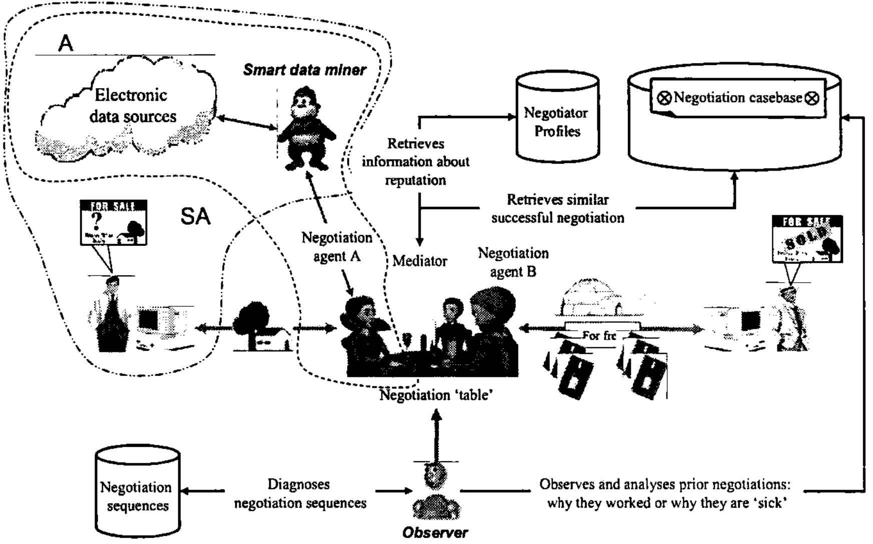


Figure 1. Current design of the curious negotiator (includes negotiation agent, mediator, observer and the smart data miner).

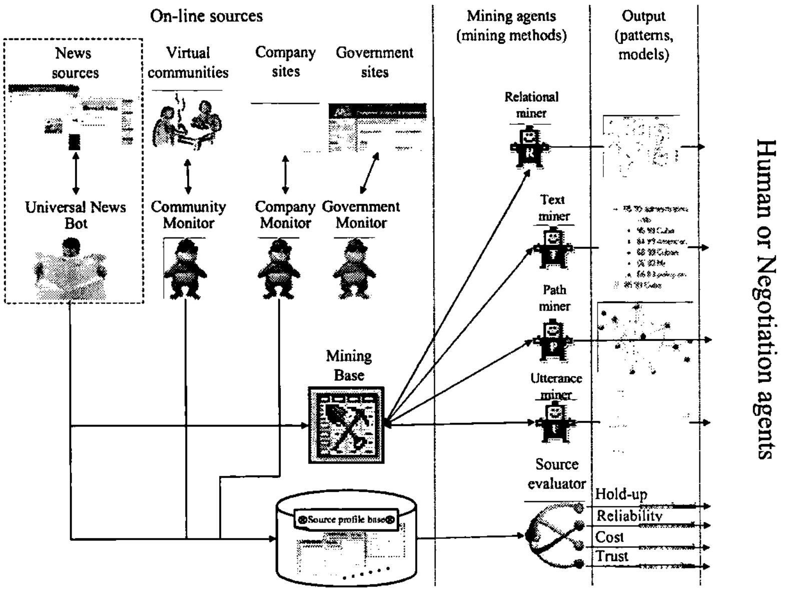


Figure 2. Smart data mining system for supporting negotiation with contextual information.

1.1 On-line News Media

Obtaining and verifying information from on-line sources takes time and resources. To reduce the impact of some delay factors on the net, the architecture of the data mining system in Figure 2 allows not only just-in-time operation, but also 'pre-fetching' some of the information that is expected to be necessary for a scheduled negotiation. In the context of news mining, the news bots fetch the news, which then are transformed into a structured form and both the structured and unstructured data are stored in the mining base (see Figure 3) for accessing by the mining agents (The fragment selected illustrated in Figure 3 shows only the text mining agent).

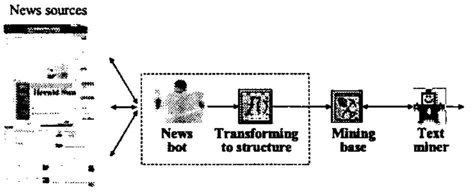


Figure 3. The news mining portion of the system

The focus of this paper is on the first phase – the automation of obtaining news from Internet sources. The news sources on the Internet include the websites of major news papers. The development of algorithms for finding the correct URLs that contain the requested news articles is within the scope of the intelligent crawler research. Major search engines, including Google (shown in Figure 4) and Yahoo recently provided a new functionality for news searching with user provided keywords. These news portals provide convenient interface for humans; answering queries in way similar to conventional search engine interface (see Figure 5).

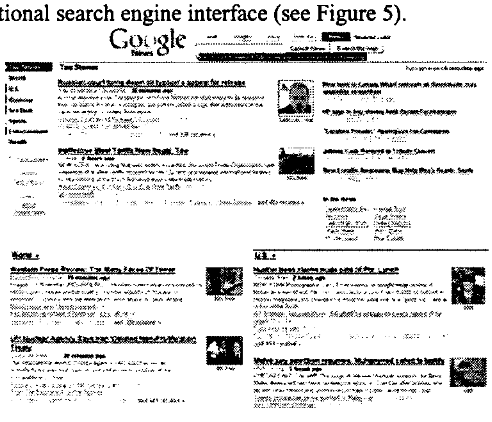
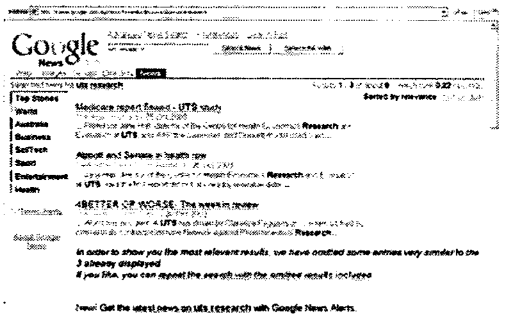


Figure 4. News search engines web interfaces



uts research

Medicare report flawed - UTS study

Abbott and Senate in health row

4BETTER OR WORSE: The week in review

In order to show you the most relevant results, we have omitted some entries very similar to the 3 already displayed.
If you like, you can repeat the search with the omitted results included.

New! Get the latest news on uts research with Google News Alerts.

Figure 5. News search engines web interfaces – response to a query.

Within the framework of curious negotiator, a generic news bot should be able to retrieve automatically, classify and store the news article obtained by a search engine in an efficient way that can be further used by the other mining agents. The initiation of the process in just-in-time mode can be by the negotiation agent (scenario "A" in Figure 1) or the human player (scenario "SA" in Figure 1). In pre-fetch mode, a source monitoring agent is subscribed to the email digests that these sources distribute [10]. These sources include 2-3 sentences news abstract and the corresponding URLs for retrieving the full articles. The trigger for fetching an article can be a negotiation scheduler, using as initial information the topic, the list of items and the description of participants.

However, the automatic retrieval by a computer program of an individual news article from the URL that is obtained either from search results or from the pre-fetched list is a tedious job since the news content can come from different web sites. Different news sources have different layout and format as illustrated by the two examples of news websites in Figure 6 and Figure 7. The layout may vary from time to time even in the news coming from the same source. Hence when automating news retrieval, even for the same news site, it is impractical to develop a static template, as it will stop working when the layout is changed. It is even more impractical (if not impossible) to develop a predefined program (template) for each news web site in the whole Internet. In this paper we present a more generic approach to retrieve news articles regardless the web site format and bringing them to the smart data mining system of the curious negotiator.

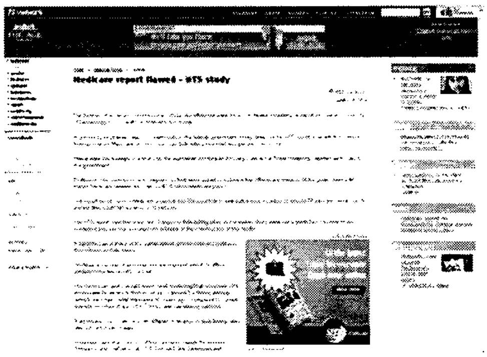


Figure 6. A version of SMH news site format

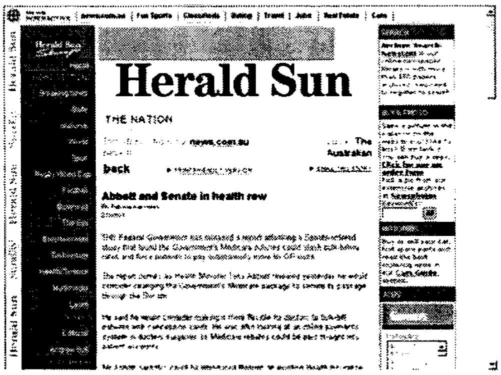


Figure 7. A version of Herald Sun news site format

2 The Universal News Bot approach

Data extraction from Web documents is usually performed by software modules called wrappers. As explained in the previous section, hard-coded wrapper by using static template is tedious, error-prone and difficult to maintain. To overcome this difficulty, significant research has been done in the area of wrapper induction, which typically applies machine learning technology to generate wrappers automatically [11, 12]. WIEN is the first wrapper induction system that defined six wrapper classes (templates) to express the structures of web sites [13, 14]. STALKER - a wrapper, more efficient than WIEN [15], treats a web page as a tree-like structure and handles information extraction hierarchically. Gao and Sterling [16] have also done significant work on knowledge-based information extraction from the internet. However, most of the earlier wrapper techniques were tailored to particular types of documents and none is specific for news content retrieval. The more recent techniques aim on data extraction from general semi-structured documents. The application of general content identification and retrieval methods to news data brings unnecessary overhead in processing. This paper proposes a technique that takes in account the characteristics of news web pages. Without loss of generality, the approach improves the processing efficiency and requires neither user specified examples or priori knowledge of the pages.

2.1 The data extraction method

The data extraction process is divided into three stages. The logical structure of the tagged (in our case, HTML) file is firstly identified and the text, which is most likely to be the news article, is extracted. During the second stage a filter is dynamically built and some extra text is filtered out if multiple documents from the same web site are available. During the third stage extracted data is validated by the developed keyword based validation method. The details are presented below.

2.1.1 Stage 1: Identifying the logical structure of the tagged file

News pages normally not only contain the news article, but more often, also related news headings, the news category, advertisements, and sometimes a search box. Although each web site may have a different format, web pages can always be broken down into content blocks. The layout in which these content blocks are arranged varies considerably across sites. The news article is expected to be the content block which is displayed on the "centre" of the page. Therefore, it is reasonable to assume that *the biggest block of text on the news web page is the news article*. Similar to McKeown et al.'s [17] approach, the biggest block of text is detected by counting the number of words in each block.

Most of web sites employ visible and invisible tables in conjunction with Cascading Style Sheets (CSS) to arrange their logical structures by using HTML table tags [18]. Table is designed to organize data into logical rows and columns. A table is enclosed within the <table></table> tag. Nested tables are normally used to form a complex layout structure. It is common for news web sites to display advertisements within news articles to attract reader's attention. This is normally done by inserting nested tables that contain advertisements and other contents in the table that contains the news article. The pseudo code of the process is presented in Figure 8.

```
Input: HTML file
Output: The largest body of text contained in a table
Begin
  1. Break down the HTML file into a one dimensional array, where
     each cell contains a line of text or an HTML tag
  2. Remove the HTML tags except <table> and </table>
  3. Set table_counter to 0
  4. For each cell in the array:
       a. if <table> tag is encountered, increase table_counter by 1
       b. if <\table> tag is encountered, decrease table_counter by 1
       c. if it is a text element, append it to the end of
          container[table_counter]
  5. Return container[i] that contains the largest body of text by
     counting the number of words.
End
```

Figure 8. Pseudo code of the algorithm for identifying the largest text block.

2.1.2 Stage 2: Building internal filters dynamically

Although most of news web sites use tables for partitioning content blocks, there are some web sites that use other methods. Also, even for the web pages that use tables as the partition method, the table with the news article may contain a few extra lines of text at the beginning or the end of the article. Therefore, extraction accuracy can be improved by developing algorithms that do not rely on table tag information. Many web sites use templates to automatically generate pages and fill them with results of a database query, in particular, for news web sites. Hence, news under same category from same source is often with same format. When two or more web pages from same source become available, a filter can be constructed by comparing the extracted text from these pages. The filter contains the common header and tail of the text. The text is compared sentence by sentence from the beginning and the end between two files. Common sentences are regarded as part of web page template. Therefore, they should be removed from the file. The pseudo code of the process is shown in Figure 9.

Once the filter is generated, text is refined by removing the common header and tail text in the filter. Since the filter is dynamically generated, it is adjusted automatically when the web site format is changed.

```
Input: two text files from the same web site, each contains a news article
Output: a data structure contains:
        String URL
        String Header
        String Tail
  1. Remove all the html tags in the files.
  2. Break down the files into one dimensional arrays (a and b), each cell
     contains a line of text.
  3. For each cell of the array from beginning
        1. if a[i] == b[i], append a[i] at the end of Header string
        2. if a[i] != b[i], break;
  4. For each cell of the array from the end
        1. if a[i] == b[i], insert a[i] at the beginning of Tail string
        2. if a[i] != b[i], break
  5. Set the URL value to the common part of the URLs of two text file
Return the data structure that contains URL, Header and Tail.
```

Figure 9. The pseudo code of dynamic filter generation.

2.1.3 Stage 3: Keyword based validation

Incorrect and out of date URLs can cause errors in the results of data extraction. Such errors can not be identified by the data extracting methods described in the previous sections. A simple validation method based on keyword frequency is developed to validate the data retrieved by the algorithms in Figure 8 and Figure 9.

The basic assumption is that a good news title should succinctly express the article's content. Therefore, the words contained in the news title are expected to be normally among the most frequent words appearing in a news article. Consequently, the words from the news title (except the stop words, which are filtered out) are considered as keywords. For situations when the news title is not available at the time of text extraction, the words in the first paragraph of the extracted data are considered as keywords, based on the assumption that title is always placed at the beginning of an article. The extracted text is regarded as the requested news article if it satisfies the following condition:

$$\min\left(w_1 \frac{l_t}{l_m}, w_2 \frac{n_k}{t_k}, w_3 k_f \right) > th \qquad (1)$$

where:

l_t total length

l_m minimum length (predefined)

n_k number of keyword that appears in the text at least once

t_k	total number of keywords
k_f	average keyword frequency
w_1, w_2, w_3	weighting values
th	threshold value (predefined)

The first term in equation 2.1 considers the total length of the extracted text. If the text length is unreasonably short, the text is unlikely to be a news article. The second term in the equation represents the percentage of the keywords that appeared in the text. The third term in the equation stands for the average frequency of the keywords that appeared in the text. The validation value takes the minimum value of these three and then compares with a predefined threshold to validate if the extracted text is the news article.

3 News Data Set Construction

The news data set for a given specific topic that will be used as the information sources for the negotiation table is dynamically constructed from on-line news articles. In stead of simply using keyword searching, the data mining agent constructs the news data set according to the concept related to the given keywords.

Similar to using searching engine, the negotiation agent provides a phrase or several keywords to the data mining agent to define the topic of the news it requests. The data mining agent submits the query to a news searching engine. In general, large amount of searching results are returned. The data mining agent only retrieve the most relevant data evaluated by the keyword frequency and their proximity position. Based on the assumption that a concept can be represented by a set of keywords, which occur frequently inside particular collection of documents, the most frequent keywords (terms) from the retrieved data set are extracted and considered to be related to the same concept. The extracted keywords are resubmitted to the search engine. The process of query submission, data retrieval and keyword extraction is repeated until the search results start to derail from the given topic. The news articles used in this section are extracted from HTML files by the algorithms described in section 2.

3.1 Key Phrase Extraction

As it is introduced in the previous section, key phrase extraction plays an important role in the data set construction process in this project. Many studies have been conducted in the area of automatic keyword generation from text documents. Most of these methods are based on syntactic analysis using statistical co-occurrence of word types in text and vector space representation of the documents [19]. Hulth [20] suggested the quality of keywords that generated by frequency analysis was significantly improved when a domain specific thesaurus is used as a second knowledge source.

Therefore, a similar approach that employs a domain specific thesaurus in the key phrase extraction process is adopted in this project. Since the frequently used words represent the topic of a document in greater degree than less frequently used words, the frequency of the words and phrases predefined in the domain thesaurus that appear in the documents are calculated and the top ranked words or phrases are considered as the keywords.

There are many publications on automatic thesaurus construction. We applied a relative simple approach based on word frequency count. The news articles in many news web sites are organized into the categories of: World, National, Business, Science (Technology), Sport and Entertainment. To build the domain specific thesaurus, a large number of news articles under each category are collected, and each of such categories represents a domain. Figure 10 shows the steps of building the database of key phases for each category (the domain thesauruses). Word stemming problem was resolved by using a simple stemming algorithm that two words are considered to have the same stem if they have the same beginnings and their endings differ in one or two characters [21]). Stop words are not counted in each document.

```
Input: document collection of each category
Output: collection of key phases for each category
Begin
  1. define the initial number of document (ni) to be used
  2. for each category
       a. randomly select ni articles from the working category to form docu-
            ment cluster A
       b. randomly select ni/(total_number_of_category-1) articles from
            every other category to form document cluster B
       c. calculate the total frequency of keywords and key phases of cluster
            A and B
       d. generate the list of most frequent keywords (phases) of cluster A
            (list A) and cluster B (list B) according to the ranking of the fre-
            quency.
       e. remove the keywords (phases) in list A if they also appear in list B
       f. increase the number of the articles to be selected by 10%
       g. repeat from a to e to generate a new list of most frequent keywords
            (phases) and compare with the list generated by the last run. If the
            difference between two lists is smaller than a predefined threshold,
            stop the process and the latest list of most frequent keyword
            (phase) is used as the keywords to define the current working do-
            main.
End
```

Figure 10. The pseudo code of domain thesaurus construction.

In 2.c of the above process, a sequence of words is defined as a phrase if it satisfies:

$$\frac{f_{seq}}{f_{average}} > th \qquad (2)$$

where:

f_{seq} frequency of the sequence of words that appears in the same sentence in the whole cluster

$f_{average}$ average frequency of each word in the sequence.

th threshold value (predefined)

3.2 News Data Set Construction process

The news data set is constructed by repeating the news retrieval and keyword extraction process. Figure 11 shows the detail procedure of the construction process. The news data domain is determined by searching the initial provided keywords (phase) in the domain thesauruses.

```
Input: domain thesauruses, initial keywords
Output: news data set
Begin
1. determine the domain thesaurus to be used according to the initial keywords
2. search and retrieve the news articles that contains the given keywords.
3. put the retrieved news articles into the news data set container.
4. calculate the total frequency of each keyword and phrase in the domain
   thesaurus that appears in the articles just retrieved and rank them from
   most frequent to less frequent.
5. if the ranking of the initial keywords is not higher than a predefined
   threshold, return the news data set.
6. calculate the total frequency of each keyword and phrase in the domain
   thesaurus that appears in each article in the whole data set and rank them.
7. select one phrase or two keywords with the highest frequency but have
   not been used for searching  as the new keywords for next search.
8. goto step 2.
End
```

Figure 11. The pseudo code of news data set construction for a given concept

4 Experimental Results

Experiments have been conducted in two steps: first to evaluate the news extraction algorithm. Second, a news data set was constructed and manually examinated.

4.1 News extraction

The proposed methods of extracting news articles were evaluated by the experiments using some of the most popular Australian and International news web sites, which are shown in Table 1. 200 pages from each URL location were tested. The average process time for each page was 436 milliseconds on a Pentium 4 1.60 GHz computer. The notions used in the table are explained below:

- *Correct* – on average 0% error rate in the extracted text of a single web page;
- *Minor Error* – on average less than 5% error rate in the extracted text of a single web page;
- *Major Error* – on average between 5% to 30% error rates in the extracted text of a single web page;
- *Error* – on average more than 30% error rate in the extracted text of a single web page

Table 1. News sites for testing the news article extraction algorithm and the results

URL Location	**Accuracy** [without Filter]	**Accuracy** [with Filter]
www.smh.com.au/national	Minor Error	Correct
www.smh.com/business	Minor Error	Correct
www.usatoday.com/news/world	Minor Error	Minor Error (Error Rate Reduced)
www.usatoday.com/news/nation	Minor Error	Minor Error (Error Rate Reduced)
http://abcnews.go.com/sections/us	Minor Error	Correct
http://abcnews.go.com/sections/world	Minor Error	Correct
http://money.cnn.com	Correct	Correct

www.cnn.com/ALLPOLITICS/	Minor Error	Correct
www.theaustralian.news.com.au	Correct	Correct
http://news.bbc.co.uk/2/hi/business	Major Error	Minor Error
http://news.bbc.co.uk/2/hi/asia-pacific	Minor Error	Minor Error (Error Rate Reduced)
http://www.reuters.com	Correct	Correct
http://news.ft.com (Financial Times)	Correct	Correct
http://dailytelegraph.news.com.au	Minor Error	Correct
www.iht.com (International Herald Tribune)	Correct	Correct
http://www.dailytimes.com.pk	Correct	Correct
http://news.xinhuanet.com/english	Correct	Correct
http://www.abc.net.au/news	Minor Error	Correct
http://news.ninemsn.com.au	Correct	Correct

Experiment results show that news articles were mostly extracted properly except BBC News (UK). After analyzing the web pages carefully, it was found that these web pages contained more than one content blocks in the table that also contains the news article, namely, the news article only occupies one of the table cell. Therefore, more experiments were conducted on this web site by using multiple documents. Experiment results show that the accuracy rate have been increased dramatically. It is because that although the content block is not correctly classified by the first step, other content blocks in the table are also extracted, but these extra content blocks in the extracted data are removed by the filtering process at the second step.

As it is shown in Table 1, by using the dynamically generated filter, the extraction accuracy has been increased considerably. The experiment confirmed the approach, which assumes that the news article is contained in a table formatting structure, and the advertisements and other content block data are embedded in nested table structure within the news article table, works well. This layout method is commonly used in most of news web sites, which makes proposed algorithms and their implementation a practically valuable tools.

During the experiment, the threshold value for validation was set to 1. Different combinations of weighting values have been tested. Experiment results showed that the validation process is highly effective. Moreover, the experimental validation results are not sensitive to the choices of weighting values.

4.2 Constructing a news data set

An experiment was conducted to build a news data set from keywords "Interest Rate". As there are large amount of news on the internet, this experiment restricted the time frame to 1 week and news sources within Australia.

Domain thesauruses were constructed by using 500 articles from each category: World, National, Business, Science (Technology), Sport and Entertainment. After the domain thesauruses have been constructed, their data remain the same for the whole experiment.

Table 2 shows the keywords (phases) used for each new search and the number of articles retrieved. The keywords for the next search were extracted from the data in the data set that has been constructed so far instead of the data from the last search results. The search process stops when the initial keywords are no longer in the most frequent keyword list generated from the last search results.

Table 2. The keywords used for each search in a data set construction process.

Keywords (phases) Used for the Search	Number of Most Related Articles Retrieved	Most Frequent Keywords
Interest Rate	23	interest rate, housing market, bank, price, bond, finance, loan ...
Housing Market	10	housing market, finance, interest rate, price, value, bank ...
Price, Bank	30	bank, price, interest rate, share, oil, economy, stock ...
Finance, Bond	12	Bond, finance, housing market, interest rate, investor, price ...
Share, Investor	27	Share, investor, housing market, bank, price, finance, value ...

In total, the news data set contains 102 news articles. Each article in the data set was manually examinated. Their contents are all within the scope of "interest rate". Once the data set is constructed, it will be further processed and used as the information source for the negotiation agent.

5 Bringing the News to the Negotiation Table

The above described tools can be used directly in the "semi-automatic" negotiation scenario (scenario "SA" in Figure 1). In this case, the information request can be initiated either by the human participant or by the negotiation agent. In both cases the keywords for initiating the news "hunt" can be extracted out of the negotiation utterances. In the case, when the negotiation agent requests the news, the keywords are filtered automatically from the dialogue and are passed to the news extraction bots (possibly with some weights based on the relative intensity with which they occur during the negotiation). In the "SA" scenario, the body of the article together with a date/time stamp and the source identifier is sufficient, as the information is assessed by the human player. An information table that contains the retrieved news text, validity of the data, most frequent keywords and other parameters is delivered to an information aggregation agent for further processing so that the information can be used by the negotiation agent efficiently. The detailed discussion of the automated utilisation of retrieved information is beyond the scope of this paper.

5 Conclusions and Future Work

The curious negotiator is the long term work in automated negotiation systems. It will blend 'strategic negotiation sense' with 'strategic information sense' as the negotiation unfolds. This requires a system capable of providing information to the "negotiation table". The smart data mining systems that support the negotiation agents are expected to operate under time-constraints and over dynamically changing corpus of information. They will need to determine the sources of information, the confidence and validity of these sources and a way of combining extracted information (models).

In this paper, we presented a method to extract relevant news article from news web sites regardless of the format and layout of the source. The article's logical structure is firstly identified by using the table tags in the tagged files. An internal dynamic filter is built to further clean up the data. Finally, a validation method is developed to validate the retrieved data. Experiment results confirm that the overall approach and the corresponding methodology and algorithms can be applied to most of the news web sites with reasonable accuracy.

In the case when a Web page is not partitioned by table tags, proposed method relies on the availability of a second document from the same web site. Although using table for page layout is the most popular method, other content partition methods should also be implemented in the system to improve the extraction accuracy.

Though developed for the curious negotiator, proposed methods can be applied for content extraction from tagged documents in mobile phone and PDA browsing area. Mobile phone and PDA have relatively slow internet access and small display area. Therefore, presented algorithms can be applied for automatic detection and display of articles from news web sites on such devices with improved efficiency and visual effect.

References

Simoff, S. J. and J. K. Debenham: Curious negotiator. Proceedings of The Int. Conference on Cooperative Information Agents, CIA-2002, Madrid, Spain, Springer, Heidelberg (2002).

Gerding, E. H., D. D. B. van Bragt, et al.: Multi-issue negotiation processes by evolutionary simulation: validation and social extensions. Proceedings Workshop on Complex Behavior in Economics. Aix-en-Provence, France, (2000).

Gomes, A. and P. Jehiel (Forthcoming): Dynamic process of social and economic interactions: on the persistence of inefficiencies. Journal of Political Economy.

Kraus, S.: Strategic Negotiation in Multiagent Environments. Cambridge, MA, MIT Press (2001).

Ströbel, M.: Design of Roles and Protocols for Electronic Negotiations. Electronic Commerce Research Journal, Special Issue on Market Design (2001).

Milgrom, P. and R. A. Weber: Theory of Auctions with Competitive Bidding. Econometrica, **50**(5), (1982).

Watkins, M.: Breakthrough Business Negotiation-A Toolbox for Managers, Jossey-Bass (2002).

Hand, D., H. Mannila, et al.: Principles of Data Mining. Cambridge, MA, MIT Press (2001).

Franz, M., A. Ittycheriah, et al.: First Story Detection: Combining Similarity and Novelty Based Approaches. In Topic Detection and Tracking Workshop Report, (2001)

Simoff, S. J. and J. K. Debenham: Time-constrained support for decision-making in e-market environments. Proceedings of the 6th International Conference of The International Society for Decision Support Systems ISDSS'01, London, UK, (2001).

Chidlovskii, B., J. Ragetli, et al.: Automatic wrapper generation for web search engines. Proceedings of the 1st International Conference on Web-Age Information Management WAIM'00, Springer. (2000).

Freitag, D. and N. Kushmerick: Boosted wrapper induction. Proceedings of the 17th National Conference on Artificial Intelligence AAAI-2000. (2000).

Kushmerick, N. and B. Grace: The wrapper induction environment. Workshop on Software Tools for Developing Agents, AAAI-98. (1998).

Kushmerick, N.: Wrapper induction: Efficiency and expressiveness. Artificial Intelligence **118**(1-2): 15-68. (2000)

Muslea, I., S. Minton, et al.: STALKER: Learning extraction rules for semistructured, Web-based information sources. Proceedings of AAAI-98 Workshop on AI and Information Integration, Menlo Park, CA, AAAI Press. (1998).

Gao, X. and L. Sterling: Semi-structured Data Extraction from Heterogeneous Sources. In T. Bratjevik D. Schwartz, M. Divitini, editor,Internet-based Knowledge Management and Organizational Memories, pages 83--102. Idea Group Publishing. (2000).

McKeown, K. R., R. Barzilay, et al.: Columbia multi-document summarization: Approach and evaluation. Proceedings of the Workshop on Text Summarization, ACM SIGIR Conference, DARPA/NIST Document Understanding Conferences (DUC). (2001).

Lin, S. H. and J. M. Ho: Discovering informative content blocks from Web documents. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD2002, ACM Press. (2002).

Salton, G.: Automatic Text Processing:The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley, (1989).

Hulth, A., J. Karlgren, A. Jonsson, H. Boström and L. Asker: Automatic Keyword Extraction Using Domain Knowledge, Proceedings of Second International Conference on Computational Linguistics and Intelligent Text Processing. *(CICLing 2001).* Mexico City, February 2001. LNCS 2004, Springer.

Andrade M, and A. Valencia: Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families, Bioinformatics (14) 600-607, (1998).

Proceedings of the 3rd Australasian Data Mining Conference – AusDM04, 6 – 7th, December, 2004, Canberra, Australia, in conjunction with the 17th Australian Joint Conference on Artificial Intelligence AI2004 and the 7th Asia Pacific Conference on Complex Systems COMPLEX 2004.

S. J. Simoff and G. J. Williams (eds).

Conference Web Site:
http://www.togaware.com/ausdm04/

Published by the University of Technology Sydney

ISBN 0-646-44379-8