Towards Semantic Web News Sites

Mostafa E. Saleh* and Ahmed Ezz*
King Abdulaziz University, Jeddah, SAUDI ARABIA

Abstract

This paper presents a model for building a Semantic News Site (SNS) based on available legacy valuable documents. It consists of four phases: extract the semantic structure for news site; build a unified semantic structure for the group of similar news sites; extract the metadata for each article in the site; and extract the knowledge from the raw text based on guiding ontology. This will enhance searching facilities, and will enable computers to understand and process documents. Consequently, this will save time consumed to extract valuable data from unstructured documents such as HTML documents. This work also, may be used to build the user profile (personalization) to extract only the required parts of the site that match the user interests. The proposed model is the first step towards the future semantic news sites. It has loosely coupled modules, and can be implemented with any distributed and heterogeneous technology such as component, agent, and web services.

Key Words: Ontology, migrator, wrapper, mediator, semantic query.

1 Introduction

The World Wide Web (WWW) gives unprecedented and unexpected revolution in the availability of electronically accessible information. Currently there are around three billion static documents in the WWW that are used by more than 200 million users internationally, and this number is growing exponentially [1, 4, 8]. On the other hand, this success results in great difficulty, and it may be impossible in the future, however, to find, to access, to present, and to maintain information of use to a wide variety of users.

The aforementioned problem has two reasons. First, the majority of pages on the Web use HTML which is the dominant presentation language for the web [4, 8]. HTML does not separate the metadata from the actual data. So there is no attention to the meaning of the included data except using the “META” tag in HTML. Secondly, the information content, however, is mainly presented in human natural language and in the editor accent [8]. Thus, the machine cannot understand that type of documents. So, it cannot make reasoning within the available data. As a result, this lends to a wide gap between the information available for tools and machines to use in creating, understanding and maintaining Web pages and the information kept in human readable form on those pages (not machine readable form). These two reasons cause serious problems in accessing and processing the available information in many applications such as: Searching for information, Presenting information, and Electronic commerce [7, 8].

Keyword searching is sometimes more effective than browsing but often returns vast amounts of data, far beyond what the user can handle. That is true if the documents found are unstructured, and the data is not organized in a machine processable form [6].

For example, to ask about the effect of “USD” on the “Gold” price, let us try the following search on “Al-Ahram” site: “تأثير الدولار على أسعار الذهب” as in Figure 1.

Although the answer is available implicitly within the resources, the result is that “No result matched” or “لا توجد مقالات تتوافق مع البحث المطلوب”. So we are in need to check the search about “Gold prices” or “سعر الذهب” then browse each document, and finally record the information from the related articles.

Data is the first-class citizen in our current and future applications [4]. We need to formulate data within available sites to be in knowledge representation form, hence answer queries such as:

- What is the exchange rate change for the USD against the Euro in the past five years?
- List all books issued about software in the last three years that discuss knowledge representation.
- What is the name of the Egyptian President?

All the aforementioned problems can be solved by the semantic web. The Web inventor, Tim Berners-Lee, defines the Semantic Web as follows: “Semantic Web is a web of data that can be processed directly or indirectly by machines” [2]. Also, “The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation” [3].
Semantic Web is about adding formal structure and semantics (metadata and knowledge) to web content for the purpose of more efficient management and access. The semantic web realization depends on the availability of a huge mass of metadata for the web contents. That metadata should be associated with the respective knowledge in the domain of interest [10]. The cornerstone of the semantic web is the ontology. Ontologies can be viewed as “schemas for knowledge bases” [16]. The term ontology came to refer to a wide range of formal representations from taxonomies and hierarchical terminology vocabularies to detailed logical theories describing a domain [16, 17].

In literature we did not find a unified algorithm to address the whole problem, but there are some related researches. Mukherjee et al. [13] introduced an algorithm for automatically dividing the HTML documents into tree-like semantic structure of the document. They depend on specific tags locations in the HTML documents, and using some heuristics to get the structure of the file.

Embley et al. [6] proposed a conceptual-modeling approach to extract and structure data from the Web. The approach describes automatic wrapper generation process for Web documents that are rich in data, narrow in ontological breadth, and have multiple records on a single page. They used unstructured documents that typically include several constants such as name, age, death date, and birth date of the decedent; a funeral date, time, and address; viewing and interment dates, times, and addresses; and names of related people and family relationships.

Thirunarayan et al. [29] described semi-automatic coarse-grain content extraction from specs. In their technique the input spec is a heterogeneous document in MS Word format. The domain knowledge and ontology is implicitly given through the domain library. The extraction output is based on SDR but can be in a number of different formats. The heterogeneous document is firstly converted into a text document preserving the text, the tables, and references to images. XML technology is used for structuring and semantic markup of the text document and for rendering the resulting XML annotated spec into different output formats using XSLT.

These researches are either concentrating on the structure extraction, or on the semantic knowledge extraction from dedicated text format. This paper presents a novel model for building a Semantic News Site (SNS) based on available legacy valuable documents. It consists of five phases: extract the semantic structure for news site which maps the semantic item with its location in the news site; build a unified semantic structure for the group of similar news sites; extract the metadata for each specific article in the site; extract the knowledge from the raw text based on guiding ontology, and user interface.

The rest of this paper is organized as follows: Section 2 presents the research motivation and the related work. Section 3 is directed to summarizing the model. Section 4 is dedicated to the extraction of the semantic structure of the news site, whereas Section 5 is specified for the unified semantic structure extraction for a similar news sites. Section 6 presents the article metadata extraction phase, and Section 7 presents the data mining module for knowledge extraction. Finally Section 8 is directed to the conclusion and future work.

2 The Proposed Model

Figure 2 shows the Semantic News Site (SNS) model. The SNS consists of five subsystems: Semantic structure extraction, Unified structural schema extraction, Article metadata extraction, News semantic data extraction, and User Interface. The SNS uses the available news sites (such as New York Times, Al-Ahram, etc.) and builds wrappers for these sites with their semantic structures. After that, it groups similar news sites that are issued in specific natural language and in specific domain, and then extracts the unified semantic structure for them. The next step is to extract the metadata for the article as a whole to enable queering based on the article level. Finally, it also extracts the news semantic data itself based on certain domain ontology and according to predefined schema.
The model uses certain ontology with each stage to guide the adaptation process or to guide the extraction of knowledge from raw text. At start, there is an initial ontology with the different categories in news sites in general such as economics, sports, education, etc. as in Figure 3. After extracting the semantic structure of each individual news site, we can update that initial ontology with the new concepts according to each news site case, for example, to add a

Figure 2: The semantic news system

Figure 3: An ontology for the news domain [12]
concept such as “books” and “culture” if they aren’t in the initial ontology. In the article metadata extraction phases, we can use metadata keywords such as “publisher”, “date”, “issue-number”, “year”, etc. Finally, with the text mining phase, we should use domain ontology concepts such as in economic domain “Bank”, “Currency”, “Metal” and relations (properties) such as “Contain”, “Affect-on”. So, using the ontological statement “Bank-WorkOn-Metal”, we can mention the instance “City Bank – WorkOn – Gold”.

Users can interact with the model in four different interaction methods. The first is the normal browsing, where users interact directly with the news sites in their normal HTML format. The second is the personalized navigation. This can be achieved after unified structure extraction, where the system can navigate to the preferred locations according to user profile. For example, to navigate to economic articles directly. The third is the querying based on article base. This is achieved after the article metadata extraction phase where all information about the article is available such as author, date, publisher, classification, title, etc. The fourth is the querying based on semantic level. This is done after extracting the knowledge from the text and representing it in suitable form of knowledge representations such as RDF. This is the most important service from the system as it saves human time and cost in formulating and carrying out such operations.

3 Semantic Structure Extraction

Our approach to semantic structure extraction is based on a simple but useful observation of Mukherjee et al. [14] where content rich web sites that update frequently such as news portals, education and e-commerce have constant semantic structure. This is because these pages are generated using populated templates from back end database. Also, the majority of Web sites are maintained by content management software and thus a large number of Web pages are machine-generated via templates. These Web pages have a fixed implicit “schema” and what changes is the content. Informally a schema for a Web page represents some concepts and relationships among them in a tree structure [22]. For example, Figure 4 is a screen shot of the Al-Ahram newspaper front page (http://www.ahram.org.eg/). Figure 4 also displayed the different semantic categories in the site such as “Major Heading news”, “Detailed News”, “News Taxonomy” and “Minor Heading News”. By applying the news ontology of Figure 3 on Al-Ahram news issue, the hyperlinks within the site can be grouped into the basic categories listed in Table 1.

Studying semantic partitioning gives several important and powerful implications in practice. First, according to [22], knowing the schema is the first step in transforming legacy HTML documents into more semantics-oriented document formats such as RDF, DAML+OIL, or OWL. Second, it eases the task of formulating queries to retrieve data from Web documents. In the New York Times example, one can pose a query to retrieve all the links under the “SPORTS” item in the taxonomy. Third, it enhances the application of audio-browsable Web content. By putting a dialog interface to the content of a Web page which is reorganized based on the knowledge of its schema, a user can easily browse its content using audio. More generally a Web site itself can be navigated using voice commands. Audio browsable Web content can significantly expand the reach of the Web to

![Figure 4: The different categories found in Al-Ahram news site](image-url)
Table 1: The groups in the different news site areas

<table>
<thead>
<tr>
<th>Group type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major Heading News</td>
<td>&quot;قرارات جمهورية بإحالة 4 مشروعات...&quot; &quot;تقريروان دوليان يشهدان بالنتائج الإيجابية&quot;</td>
</tr>
<tr>
<td>Minor Heading News</td>
<td>مبارك والبشر يشهدان اليوم توقيع وثيقة القاهرة للوفاق السوداني &quot;مع بداية جولتهما في المنطقة...&quot;</td>
</tr>
<tr>
<td>Detailed News</td>
<td>&quot;قرارات جمهورية بإحالة 4 مشروعات...&quot; &quot;تقريروان دوليان يشهدان بالنتائج الإيجابية...&quot;</td>
</tr>
<tr>
<td>News Taxonomy (Categories)</td>
<td>&quot;الصفحة الأولى&quot; &quot;مصر&quot; &quot;الوطن العربي&quot;</td>
</tr>
</tbody>
</table>

visually challenged individuals. Fourth, according to [15], browsing based on the logical structure can be integrated with summarization to yield still greater reductions in information overload. Finally, semantic partitioning can enable the creation of self-repairing wrappers, the technology that provides a database-like interface to Web documents [22].

The news site taxonomic structure extraction process is shown in Figure 5. The extract issue groups module extracts all the hyperlinks in the issue and divides them into related groups according to spatial locality. Spatial locality [12, 22] can be identified by looking for similarities in the path structures of the corresponding DOM trees. This is used to check that a certain element belongs to a certain group of links or not. For example, the node-to-root path for all links under the same group will have the same node-to-root path, and will share the same visual characteristics.

ExtractGroups(file f) Algorithm

Input: HTML file that includes the Web page for the news site

Begin

Calculate n, the number of hyperlink elements <A href…> in the document
Define an array hyperlink[n], paths[n], groups[n]
Define groups=0, hyperlink=0
Put 0 in all elements of groups array
For each hyperlink element <A href….>
  Calculate the node-to-root path // The distance to the root element.
  Store the hyperlink in array hyperlink
  Store the path in an array paths[ hyperlink]
  Increment hyperlink

For each element in the paths[ ] array
  If paths[i]=paths[i+1] and paths[i] is not included in a group (group[i]=0)
    Increment groups
    groups[i]=groups
    groups[i+1]=groups
  If paths[i]=paths[i+1] and paths[i] is included in a group (group[i]>0)
    groups[i+1]=groups

End

This algorithm extracts all the hyperlink elements in the document along with their paths to root. The path to root is calculated by tracing the index of HTML elements “table”, “tr”, and “td”. For example, in Figure 6, “T[3] R[1] D [1]” means that this hyperlink includes the third table, the first row, and the first cell. If two or more hyperlinks share the same table, and the same row, this means that they are in the same distance to root; hence they are within the same group.

After applying “ExtractGroups” algorithm, we can find some groups that represent “Major Heading News”, “Detailed News”, “News Taxonomy”, and “Minor Heading News”. By comparing items in each group in two different issues of the same news site, we can conclude that “Major Heading News”, “Minor Heading News”, and “Detailed News” groups have different items from day to day, but at the same time, all the
other template groups such as “News Taxonomy” and the other advertising hyperlinks are fixed in each issue of the news site.

In automatic news taxonomy matching, the news ontology concepts are matched with the hyperlink active elements to find if they match with the ontology depicted in Figure 3. In some cases we can find similarity (Synonymous keywords) that can match with the required concept. Table 2 can be used as an initial guide for synonymous concepts in the “News taxonomy”. Table 2: Synonymous keywords matched with categories of Figure 1

<table>
<thead>
<tr>
<th>Ontology Concept</th>
<th>Mapping keywords (Synonymous)</th>
</tr>
</thead>
<tbody>
<tr>
<td>National</td>
<td>National, Egypt, U.S., France…</td>
</tr>
<tr>
<td>International</td>
<td>International, World</td>
</tr>
<tr>
<td>Business</td>
<td>Business, Finance</td>
</tr>
<tr>
<td>Sports</td>
<td>Sports, Baseball, Football…</td>
</tr>
<tr>
<td>Health</td>
<td>Health, Fitness</td>
</tr>
<tr>
<td>Economic</td>
<td>Economic, Money</td>
</tr>
<tr>
<td>Arts</td>
<td>Arts, Movies</td>
</tr>
<tr>
<td>Culture</td>
<td>Culture, Books, Readings</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Science, Technology</td>
</tr>
</tbody>
</table>

In Arabic news sites, the majority of them use images in their hyperlinks to give the best appearance in their presentation. So if the model did not succeed in extracting the semantic structure automatically, manual step is used to match the concepts with the hyperlinks in the document. The news site semantic structure can be in a tabular form or in a tree structure to give different levels of details. Whatever it is, the structure (node) takes the form of:

1) Subject: The topic of the content of the news
2) Link: The site address of the subject in that newspaper.

For example, the semantic structure for Al-Ahram has the structure in Figure 7a, and its RDF representation is in Figure 7b. RDF is used to represent meanings of certain domain and it is also used to represent metadata. Meaning is expressed in RDF as encoded triples [9, 19, 21] which make assertions that particular resources (subjects) have properties (predicates) with certain values (objects). A resource may be an entire Web page, a part of a Web page, a collection of pages (for example, an entire Web site), and another RDF triple, or even non-electronic media such as a printed book. A property is a specific aspect, characteristic, attribute, or relation used to describe a resource.

4 Unified Schema Extraction

The unified similar news sites schema or semantic structure for a group of related sites is very important for some reasons. First, it enables the isolation of the user from directing the query engine to certain addresses that he must know himself. So, it constructs like a general interface to the underlying sites and directs the engine internally to the specific location managed by the system. Second, it enables in the future the interoperability between the sites which is one of the most challenging issues. Third, it helps in personalizing the browser according to customer profile. As it is independent of syntactic structure, it can monitor the structure change if one of the main subjects is changed. To build the unified schema, we will use the guiding ontology after updating it if required to guide in the process of extracting the unified (global) ontology. To get the unified schema, we apply the following algorithm:

**Algorithm ExtractUnifiedSchema**

Input the group of related sites schemas
Begin

Define a two dimensional array unifiedSchema which has rows n equals to the number of the related news sites to be matched, and the cols m equals to the number of available concepts in the guiding ontology
For each site ontology
    For each concept in the site ontology
        If the concept matches with the current concept in the guiding ontology then
            update the unifiedSchema [n][m] with the hyperlink of the concept
        Else
            Move to the next concept
End

The unified schema has three items:

1) Subject: The topic of the concept in the news category.
2) News Site Name.
3) Link: The link to that concept in the news site.

Table 3 shows the structure of the unified semantic structure. The first row contains the names of the news sites. The first column contains the names of the subjects (concept category) economic, sports, education, etc. Note that, the link may be empty if the news site does not cover this subject.
5 Article Metadata Extraction

The unified structure is just the wrapper for the SNS to enable navigation to the required title (Semantic concept) in the news site. We will go further in depth by processing the article itself by extracting the related metadata. According to [11], metadata is “information that makes data useful”. Recently metadata have been used to describe digital resources available across networks. In this context, the metadata concept plays an important role, not only because it allows a description of available resources, but also because it supports a variety of functions that users have come to expect from search mechanisms, such as: data localization, data assessment, data selection, and data retrieval [11].

Table 3a: The unified schema structure

<table>
<thead>
<tr>
<th>Subject(n)</th>
<th>Link(1)</th>
<th>Link(2)</th>
<th>Link(3)</th>
<th>……</th>
</tr>
</thead>
<tbody>
<tr>
<td>News Site(1)</td>
<td>News Site(2)</td>
<td>……</td>
<td>News Site(n)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3b: Portion of the unified schema for two news sites

<table>
<thead>
<tr>
<th>Category</th>
<th>News Site(1)</th>
<th>News Site(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>National</td>
<td>…/Index.asp?CurFN=egyp0.htm&amp;DID=8862</td>
<td>…/Index.asp?CurFN=worl0.htm&amp;DID=8862</td>
</tr>
<tr>
<td>International</td>
<td>…/Index.asp?CurFN=arab0.htm&amp;DID=8862</td>
<td>No link</td>
</tr>
<tr>
<td>Business</td>
<td>…/Index.asp?CurFN=econ0.htm&amp;DID=8862</td>
<td>No link</td>
</tr>
<tr>
<td>Sports</td>
<td>…/Index.asp?CurFN=sport0.htm&amp;DID=8862</td>
<td>…/Index.asp?CurFN=sporo.htm&amp;DID=8862</td>
</tr>
<tr>
<td>Health</td>
<td>No link</td>
<td>No link</td>
</tr>
</tbody>
</table>
There are three major aspects for the deployment of metadata: resource description, metadata production, and metadata use. The first aspect concerns the information that will be expressed by the metadata. The resource type and purpose of descriptors will determine this major question. The second aspect is production of metadata. Metadata represent a summary of data descriptions. When manually generated, it constitutes an expensive process. There is a general tendency to make this generation an automatic procedure, whenever possible. The third aspect concerns metadata use and access. It is especially important as a mechanism for resource location in distributed networks environments like the Internet. It contains embedded information like resource identification, subject and structure description, etc., all important for the resource access and availability [15].

To describe metadata, we can use Dublin Core metadata [5]. The Dublin Core Metadata Element Set is a simple vocabulary (which can be encoded in RDF) for describing fundamental properties of Web resources such as the creator (“The person or organization primarily responsible for creating the intellectual content of the resource”) and date (“A date associated with an event in the life cycle of the resource”). Also, the Publishing Requirements for Industry Standard Metadata (PRISM) specification defines an XML metadata vocabulary for syndicating, aggregating, post-processing and multi-purposing magazine, news, catalog, book, and mainstream journal content. PRISM provides a framework for the interchange and preservation of content and metadata, a collection of elements to describe that content, and a set of controlled vocabularies listing the values for those elements [18]. Compatibility with these standards is highly desirable because using such metadata agencies can make their contents more accessible to other applications, which are also compatible with them.

The article metadata is collected from many sources:

- **Date** Date of news.
- **Type** Category of the article.
- **Publisher** The site publisher.
- **Copyright** The copyright of site issues.

The article itself gives us:

- **Author** The writer of the article.
- **Title** The title of the article.
- **Format** The format used to produce the article.
- **Word Count** The count of the number of words in the article.
- **Text** The content of the article.

The whole site issue metadata can be represented in a form similar to Figure 8, and the article metadata is represented in Figure 9.

### 6 Article Text Mining

This is the last stage in preparing the knowledge repositories for the user query and question answering. It includes extracting the valuable knowledge within the article and storing it in the knowledge base. This stage is considered as a migration stage where the raw article text is migrated into structured knowledge that can be processed by the machine. This stage needs two types of repositories. The ontology repository includes the main categories in the domain of interest such as economic, sports, education, etc. Figure 10 shows sample ontology in the economic field. The second repository is for the instances that match with the ontology. It represents the knowledge base that should be updated by more documents issued by the sites, and also, it will be inferred when answering a query or question.

By studying the text in Figure 11, we can conclude the following relations:

- Gold is a metal.
- LE is a currency.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:rdf xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/" >
<rdf:Description rdf:about="http://www.ahram.org.eg/>
<dc:description>Al-Ahram News Site</dc:description>
<dc:date>2005-6-1</dc:date>
<prism:copyright>&copy; Copyright 2001, Al-Ahram, All rights reserved.</prism:copyright>
</rdf:Description>
</rdf:RDF>
```

Figure 8: Issue metadata represented in RDF
The affect may be [win| lose]

The relation of Gold with respect to LE is decreased. Any relation in any domain of interest can be formulated in triple form. The first item is the resource to be described (subject of the relation), the second is the relation type (verb of the relation), and finally the value of the property (object of the relation) [9, 19, 21]. For example “metals” “affect [win, loss, no-effect]” “currencies”, and “human” “eat” “food”. To formally mention that "Gold wins dollar" we can use the ontology in Figure 12.

To formulate the relation between the Gold metal and the LE currency, we can use Figure 13. It explains that the gold price is decreased with respect to the LE, with a piece of 72 LE. By studying the relations in the domain of interest, we can build the complete ontology for that domain and then we
can use it in extracting the knowledge hidden within the text.

7 Conclusion And Future Work

The realization of the semantic web needs much more effort. An important milestone in this realization is to put a notion for the applications that use metadata explicitly in describing related concepts, and to clearly describe the relation of the concepts with each other through ontologies. This paper has presented the notion of a semantic model for SNS to benefit from the daily huge amount of documents published for news sites. Current news sites aim to just publish data for user consumption not for machine to process. In this work, we identified the need to wrap HTML documents with RDF. We argue that instead of designing dedicated programs to handle each kind of HTML news document for each specific application, we could design a generic architecture for handling a majority of news documents.

As to future directions, we intend to fully implement that model to help in solving the problem of data overloading in an important source of the problem which is the news field. Also, building tools for automation of the whole process is an important issue in the future. Finally, we plan for the adaptation of the model to fit with other fields of our life such as e-learning, e-business, etc.

References


MOSTAFA E. SALEH is an Associate Professor in the Computer Science Department, King Abdulaziz University, Jeddah, Saudi Arabia. He has a Ph.D. in Computer Engineering from Mansoura University, Egypt, 2000. He has published articles in areas of Document Processing, Semantic Web, and E-Learning. He is also involved in many academic projects in the E-Learning area.

AHMED EZZ (photo not available) is an Assistant Professor in the Computer Science Department, King Abdulaziz University, Jeddah, Saudi Arabia. He finished his Ph.D. in Computer Engineering at Al-Azhar University, Egypt. He has published articles in areas of Software Agents, Semantic Web, and E-Learning.