A Multi-Agent System for Web Document Authoring

M. Pérez-Coutiño, A. López-López,

M. Montes-y-Gómez and L. Villaseñor-Pineda

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
Luís Enrique Erro No. 1, Sta Ma Tonantzintla, 72840, Puebla, Pue, México.
{mapco,allopez,mmontesg,villasen}@inaoep.mx

Abstract. Current efforts on the semantic web are mainly focused on the creation of recommendations and standards for adding semantic descriptions to web resources. This situation represents a huge challenge to content creators that have to construct manually such descriptions, implying high costs in material and human resources. This paper presents a multi-agent system that automates partially this task, i.e. the authoring of web documents, reducing content creators labor. This system automatically extracts descriptive information from a set of documents in Spanish language, and constructs two output (web) document collections from them. The first collection is a set of meta-information descriptions based on the Dublin Core specifications. The second output is a collection of XHTML documents for human visualizing and browsing. In order to build the two output collections, the proposed multi-agent system applies several intelligent text processing approaches. This paper describes these approaches, as well as, the methodology used to encode the extracted metadata. It also reports results from processing three document collections of about 45 MB of text, including their associated resources –descriptions and hypertext– generated by the system.

Keywords: Semantic web, web document authoring, metadata extraction, automatic link generation, multi-agent system.

1 Introduction

The Internet has become the preferred media for interchange of both information and knowledge. However, nowadays, this information is mainly designed for human usage and not for the computers (Berners-Lee et al., 2001). Several problems arise as a result of the unstructured nature of the web information. For instance, the information retrieval engines are incapable of getting appropriate results, with acceptable levels of both recall and precision (Kobayashi and Takeda, 2000); and automated approaches such as information software agents cannot reach their goals (Berners-Lee et al., 2001).

In order to improve and extend the automated usage of the web, its information must be enriched. A common way to enrich this information is to include *meta-information*, i.e. information about the resource itself describing its content and its

relations to other resources, in a meaningful way to machines. Currently, humans, with expertise in a specific domain, construct the meta-information descriptions.

There are several initiatives focused on creating standard schemas to capture semantics of many domains (Egnor and Lord, 2000). For instance, the World Wide Web Consortium (W3C) promotes the *semantic web* initiative with the aim of extending the current web to facilitate web automation and universally accessible content, and the Dublin Core Metadata Initiative (DCMI) considers the development of interoperable online metadata standards that support a broad range of purposes and business models.

Despite of these standardization efforts, it is clear that the consolidation of the semantic web requires the creation of automatic methods for both, authoring and retrieval tasks.

This paper focuses on *information authoring*. It proposes a multi-agent system for automatically extracting descriptive information from Spanish documents, and constructing their meta-information descriptions based on the Dublin Core element set. Additionally, this system generates a collection of XHTML documents allowing effortless visualization of the extracted information.

The proposed authoring system is expected to be of impact since the extracted meta-information can be used to improve web processing by search engines and web agents.

The rest of the paper is organized as follows. Section 2 reviews previous work on metadata creation. Section 3 presents an overview of our multi-agent system. Section 4 shows some experimental results obtained from the processing of three document collections of about 45 MB of text. Finally, section 5 discusses the main contributions and further improvements.

2 Related Work

The efforts on the creation of *metadata schemes* for the web are lead by the World Wide Web Consortium (W3C) through its semantic web initiative[1]. They define the semantic web as an abstract representation of data on the web, mainly based on the RDF standard (SemanticWeb, 2001).

Other domain specific communities, such as the *Dublin Core Metadata Initiative*[2] (DCMI), also use RDF/XML for publishing data on the web. Additionally, there are several metadata initiatives for encoding bibliographic resources and literary and linguistic texts. Two popular examples are the Machine Readable Cataloging Record (MARC), and the Text Encoding Initiative (TEI).

Our work is based on the DCMI because of its simplicity, semantic interoperability, international consensus and extensibility. The entire DCMI considers 15 elements grouped in 3 categories: content, intellectual property and instantiation. The proposed system considers all three categories, but it mainly focuses on the extraction of the subject and relation elements.

The task of *subject extraction* is typical for information retrieval systems (Baeza-Yates and Ribeiro-Neto, 1999). Traditional systems use statistical methods to select

[1] www.w3.org/2001/sw/

[2] www.dublincore.org

the set of words that best represents the content of the documents. On the contrary, recent approaches tend to apply simple natural language methods to obtain representative phrases as document descriptions (Strzalkowski et al., 1997; Buckley et al., 1995).

Our system represents document content as a list of topics, i.e. sequences of words indicating a unique entity. Currently, we extract these topics based on simple heuristics for Spanish language; however we are developing a Mexican-Spanish POS tagger for a further richer analysis[3].

Most work on relation extraction, also known as *automatic link generation*, considers applications not necessarily working on Internet or the Web. For instance, Allan (1996) automatically generates links for a set of documents based on their similarity measure; Golovchinsky (1999) presents the system VOIR capable of identifying candidate links based on some user-specified topics; and Kaindl and Kramer (1999) propose a semiautomatic glossary link generator allowing the interaction with the users.

Our system is based on Allan's work, but employs improved document representation (the list of topics); produces metadata in the RDF/XML format in accordance with the DCMI and semantic web guidelines; and generates a XHTML output collection that becomes a hypertext volume for human reading and navigation.

3 The Multi-Agent System

As we mentioned, the main goal of our multi-agent system (MAS) is to automate the authoring of metadata and hypertext for large collections of electronic Spanish documents. In order to reach its goal, the MAS carry out two main tasks:

- *Input processing*, consisting of the identification of language, subject, and other attributes for each document, as well as the inter-document relations.
- *Output generation*, considering the creation of a set of metadata in the RDF/XML format suited for machine processing, and a collection of XHTML documents for human reading and browsing.

The general architecture of our MAS is shown in the figure 1. It consists of two layers of agents: the input processing and the output generation agents. These agents are heterogeneous and extensible, i.e., each agent has a specific plan, and its behavior is separated from its functionality. This design allows easily implementing and modifying the agent algorithms.

The MAS was developed based on the Jack Intelligent Agent Framework (Busetta et al., 1999). This framework provides a set of Java components for developing multi-agent systems in accordance with the Belief-Desire-Intention model. The communication among the agents is based on the contract net protocol.

The subsequent sections describe the goals and functionality of the four main agents: the subject extractor, the relation finder, the metadata generator and the XHTML generator.

---

[3] "Etiquetador de Partes de la Oración para el Español de México", Project CONACYT R31886-A, 1999-2002.

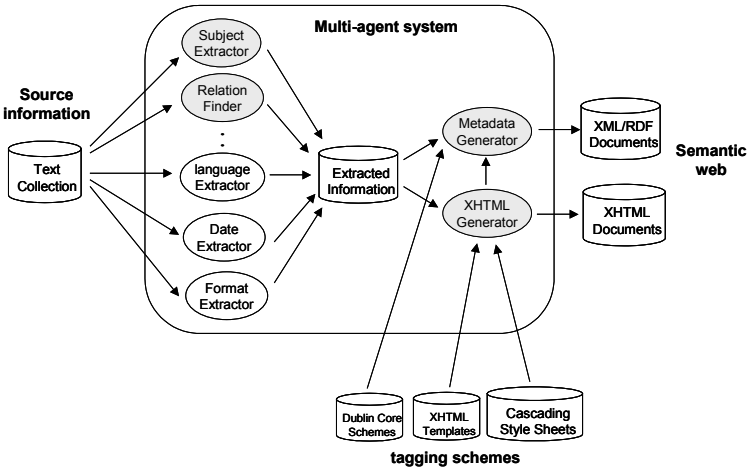

Figure 1. General architecture of the multi-agent system

3.1 Subject Extractor

The subject extractor agent has two main tasks: to identify the candidate topics for each document of a given collection, and to build a formal representation of their content.

In order to identify the set of topics of a document, this agent uses a method similar to that proposed by Gay and Croft (1990), where the topics are related to noun strings. Basically, this agent applies a set of heuristic rules specific for Spanish, based on the proximity of words, that allows identifying and extracting key phrases. These rules are driven by the occurrence of articles and the preposition *de* ('of') along with nouns or proper names. Some morphological inflection patterns (typical endings of nouns and verbs) are also taken into account. For instance, given the following paragraph, the subject extractor agent selects the underlined words as the candidate topics:

"*Góngora Pimentel* aseguró que estas *demandas* se resolverán en un *plazo* no mayor de 30 días y que sin duda la *demanda interpuesta* por el *PRD* ante la *Suprema Corte de Justicia* se anexará a la que presentó el *Partido Acción Nacional*".[4]

Then, based on the candidate topics, this agent builds an enriched representation of the documents. This representation is expressed as a weighted vector of topics in a given *n*-dimensional vector space. That is, for a given collection of documents $D = \{d_i\}$, with a corresponding set of topics $\{t_1, \ldots, t_n\}$, the new document representation is formally expressed as follows:

[4] '*Góngora Pimentel* confirmed that these *demands* will be satisfied in a *period* not longer than 30 days and that without any doubt the *demand introduced* by the *PRD* to the *Justice Supreme Court* will be added to that presented by the *National Action Party*.'

$$d_i \rightarrow \vec{d}_i = (w_i(t_1), w_i(t_2), \ldots, w_i(t_n)), \text{ where:}$$

$$w_i(t_j) = \frac{f_{ij}}{\sum_{k=1}^{n} f_{ik}}$$

In these formulas, $w_i(t_j)$ is the normalized weight of the topic j in the document i; f_{ij} is the number of occurrences of the topic j in the document i; and n is the number of topics in the whole collection.

3.2 Relation Finder

The goal of the relation finder agent is to identify the most significant inter-document relations. Basically, this agent finds the set of thematically related documents for each item of the given source collection.

In order to accomplish its goal, the relation finder agent computes the similarity for every pair of documents in the source collection, and then determines the most important connections.

The similarity measure used is based on the Dice coefficient:

$$s(d_i, d_j) = s_{ij} = \frac{1}{2} \sum_{\forall t \in d_i \cap d_j} w_i(t) + w_j(t)$$

Here, the topic $t \in d_i \cap d_j$ is a common topic of both documents d_i and d_j, and $w_k(t)$ indicates the weight of the topic t in the document d_k.

The criteria used to determine the set of related items associated to the document d_i, after computing the similarities is the following:

$$R_i = \{ d_j \mid s_{ij} \geq s_\mu,\ j \neq i \}, \text{ where:}$$

$$s_\mu = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=i+1 \\ s_{ij}>0}}^{N} s_{ij}$$

Here, R_i is the set of thematically related documents for the document d_i, s_{ij} is the similarity measure of documents d_i and d_j, and N is the number of documents in the whole collection. Basing this criterion on the average similarity among documents allows producing the associated set of items independent of how homogeneous is the collection. That is, even in highly heterogeneous collection (a very diverse set of topics), we can obtain existing relations.

3.3 Metadata Generator

This agent gathers the extracted metadata information (i.e., the subject and the relation elements) along with other information coming from the documents, such as last modification date, the language, and the format. Then, it encodes these elements based on the recommendations for generating Dublin Core metadata in RDF/XML (Becket and Miller, 2002). The resulting metadata set serves as machine-readable information, allowing their automatic processing by software agents and search engines.

In order to generate the metadata in the Dublin Core format, this agent uses the template detailed below. In this template, the bold-font elements are automatically generated by our system, while the rest of them are pre-configured or post-produced by the user. For instance, in our private experiments the system fills the creator element with the system signature and the publisher element with the data from our laboratory.

```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF SYSTEM "http://dublincore.org/2000/12/01-
dcmes-xml-dtd.dtd">
<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>
  <dc:creator>cre</dc:creator>
  <dc:contributor>con</dc:contributor>
  <dc:publisher>pub</dc:publisher>
  <dc:subject>sub</dc:subject>
  <dc:description>des</dc:description>
  <dc:identifier>ide</dc:identifier>
  <dc:relation>rel</dc:relation>
  <dc:source>sou</dc:source>
  <dc:rights>rig</dc:rights>
  <dc:format>for</dc:format>
  <dc:type>typ</dc:type>
  <dc:title>tit</dc:title>
  <dc:date>dat</dc:date>
  <dc:coverage>cov</dc:coverage>
  <dc:language>lan</dc:language>
</rdf:Description>
</rdf:RDF>
```

The metadata in RDF/XML format corresponding to an example document is showed in the section 4.

3.4 XHTML Generator

As we mentioned in the section 2, our system generates two different outputs for a given document source collection. One is a set of XML/RDF documents corresponding to the semantic description of the input documents (as described in section 3.3). The other output is a collection of XHTML documents. The main purpose of this collection is to become a *hypertext volume* for human reading and browsing. The output collection is based on a template that fulfills the standard XHTML 1.0 proposed by the World Wide Web Consortium (W3C), and includes the following set of metadata: title, creator, publisher, date, subject and relation. It also contains the source document and a pointer to the Dublin Core document representation instrumented by the tag: *<link rel="meta" href="SomeURL/xml/file.shtml.rdf">*.

The XHTML output corresponding to the example text is showed in the section 4.

4 Experimental Results

4.1 The test collections

In order to prove the functionality of the proposed MAS, we analyzed three document collections: *News94*, *ExcelNews* and *Nexos90s*. These collections are all in raw text format (i.e. ASCII). They differ from each other in their topics and in the document average size.

Next, we describe the main characteristics of the three test collections. More details are in table 1.

Collection News94

News94 is a set of 94 news documents. The average size per document is 3.44 Kb, and the biggest document size is 18 Kb. This collection is a subset of the ExcelNews data set.

Collection ExcelNews

This collection consists of 1,357 documents. These documents contain national and international news from 1998 to 2000 as well as cultural notes about literature, science and technology. The document average size is 3.52 Kb, and the biggest document size is 28 Kb.

An important characteristic of the ExcelNews collection is the variety of writing styles and lexical forms of its documents, causing a large distribution of terms in the vocabulary.

Collection Nexos90s

This collection contains the articles that appeared in the issues of the 1990`s from the Mexican magazine "Nexos". It includes 120 documents –one per month– with an average size of 344 Kb (i.e., approximately 100 pages). Their content is mainly political, but some other topics, such as literature, art and culture, are also treated.

Table 1. Main data of test collections

Collection	Size (Mb)	Number of documents	Average document size	Average number of pages	Number of lexical forms	Number of terms
News94	372 Kb	94	3.44 Kb	124	11,562	29,611
ExcelNews	4.81	1357	3.52 Kb	1,642	41,717	391,003
Nexos90s	41.10	120	344 Kb	14,029	133,742	3,433,654

4.2 Results

Table 2 summarizes the results obtained from the analysis of the test collections. These results consider three main aspects: (1) the topic distribution of the test collections, (2) the required time for their analysis, and (3) the connectivity level of the resulting hypertext document sets.

Table 2. Main results from the collection analysis

Collection	Topics	Instances of topics	Indexing time	Searching time	Connected documents	Relations	Average of related documents
News94	2,571	4,874	0’’.26	0’’.55	90	459	5
ExcelNews	24,298	72,983	3’’.56	3’50’’.59	1350	47,486	35
Nexos90s	145,813	1,096,421	5’14’’.72	580’26’’.7	118	3,803	32

Next, there is an example of the metadata set gathered from a given input document. Then, figure 2 shows the resulting XHTML document.

```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF SYSTEM "http://dublincore.org/2000/12/01-
dcmes-xml-dtd.dtd">
<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>
<dc:creator>AcreS, Multi-Agent System for web document authoring
</dc:creator>
<dc:publisher>Language Technologies Lab, Csc, Inaoe
</dc:publisher>
<dc:subject>Presidente Ortiz Rubio, selección de candidato, PRI,
Partido Socialista Fronterizo, PNR, Poncho Martínez Domínguez,
fuerza caciquil, Supuso Madrazo, Polo Sánchez Celis, Javier Ro-
mero, derrota automática, Madrazo, partido callista, PPS de Lom-
bardo, desaparición de poder, entrega final, venia central, Por-
tes Gil, candidato, poca política
</dc:subject>
<dc:identifier>010698-1Lunes</dc:identifier>
<dc:relation>http://ccc.inaoep.mx/~mapco/acres/n94/020598Sabado.
xhtml </dc:relation>
<dc:relation>http://ccc.inaoep.mx/~mapco/acres/n94/050698-
1Viernes.xhtml </dc:relation>
<dc:relation>http://ccc.inaoep.mx/~mapco/acres/n94/150598-
1Viernes.xhtml </dc:relation>
<dc:relation>http://ccc.inaoep.mx/~mapco/acres/n94/180698-
1Jueves.xhtml </dc:relation>
<dc:relation>http://ccc.inaoep.mx/~mapco/acres/n94/200698Sabado.
xhtml </dc:relation>
<dc:relation>http://ccc.inaoep.mx/~mapco/acres/n4/280698Domingo.
xhtml </dc:relation>
<dc:relation>http://ccc.inaoep.mx/~mapco/acres/n94/300598-
1Sabado.xhtml </dc:relation>
<dc:relation>http://ccc.inaoep.mx/~mapco/acres/news94/300698Mart
es.xhtml </dc:relation>
<dc:format>xhtml</dc:format>
<dc:date>06-01-1998</dc:date>
<dc:language>es</dc:language>
</rdf:Description> </rdf:RDF>
```

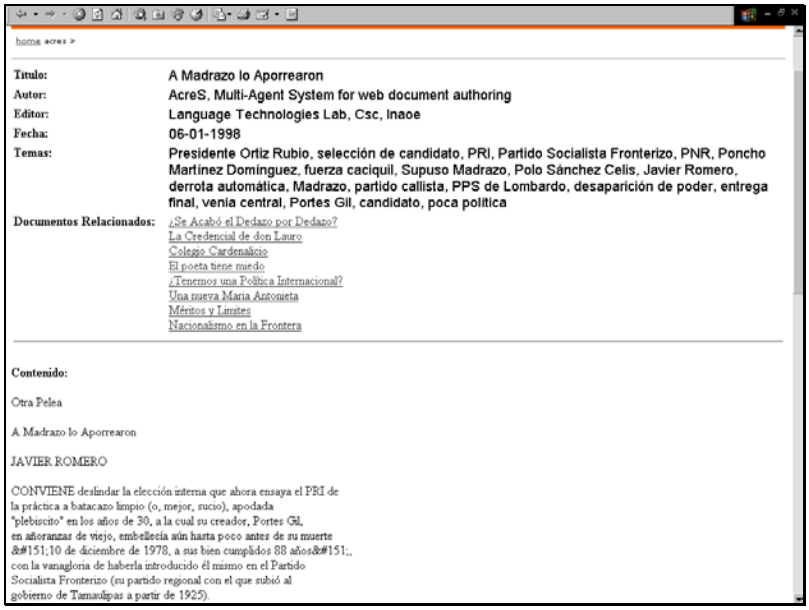

Figure 2. A sample page generated by the MAS

5 Conclusions and Future Work

We have proposed a multi-agent system that partially automates the generation of semantic descriptions for web resources, considering the identification of language, subject, date and the inter-document relations, as well as the generation of two kind of outputs, one suited for machine processing, and other for human reading and browsing.

We performed experiments with collections of different sizes and characteristics. Among the conclusions obtained from these experiments are the following:

- Representing the documents by a set of topics instead of a set of keywords speeds up document processing, reduces the number of relations among documents, and improves semantics of subject and relation metadata.
- Applying a topic weight scheme that considers their inverse frequency causes the identification of several relations with no semantic meaning since favours terms in relatively few documents, reducing considerably the influence of central topics.
- In contrast, using a criterion based on the average similarity for identifying relations among documents allows processing both homogeneous and heterogeneous document collections.

As future work we plan to: (1) apply shallow NLP techniques such as POS tagging to improve document topics identification, (2) propose an extension to the Dublin Core template in order to capture the semantics of the inter-document relations, and (3) explore some other criteria for establishing inter-document relations.

Acknowledgements. This work was done under partial support of CONACYT (project 31128-A), SNI-Mexico, and the Human Language Technologies Laboratory of INAOE.

References

1. Allan J. Automatic Hypertext Link Typing. *Proc. of ACM Conference of Hypertext 96*, Washington, D.C, 1996.
2. Baeza-Yates R., and B. Ribeiro-Neto. *Modern Information Retrieval*, Addison-Wesley, 1999.
3. Beckett D., and E. Miller. *Expressing Simple Dublin Core in RDF/XML*, Institute for Learning and Research Technology (ILRT) University of Bristol**;** W3C, 2002-07-31. URL: http://dublincore.org/documents/2002/07/31/dcmes-xml.
4. Berners-Lee T., J. Hendler and O. Lassila. The Semantic Web, *Scientific American*, May 2001.
5. Buckley C., A. Singhal, M. Mitra and G. Salton. New Retrieval Approaches using SMART: TREC 4. *Proceedings of the 3rd Text Retrieval Conference (TREC-4)*, 1995.
6. Busseta P., R. Rönnquist, A. Hodgson, and A. Lucas. *Jack Intelligent Agents – Components for Intelligent Agents in Java*, Technical Report 1, 1999.
7. Egnor D., and R. Lord. Structured Information Retrieval using XML, *ACM SIGIR 2000 Workshop On XML and Information Retrieval*, Athens, Greece, July 2000.
8. Gay L., and W. Croft. Interpreting Nominal Compounds for Information Retrieval. *Information Processing and Management* 26(1): 21-38, 1990.
9. Golovchinsky G. What the Query Told the Link: The Integration of Hypertext and Information Retrieval. *Proc. 8th ACM Conference on Hypertext*, 1997.
10. Kaindl H., and S. Kramer. Semiautomatic Generation of Glossary Links: A practical solution. *Proceedings of the tenth ACM Conference on Hypertext and Hypermedia*, Darmstadt, Germany, 1999.
11. Kobayashi M., and K. Takeda. Information Retrieval on the Web. *ACM Computing Surveys*, Vol. 32, No. 2, p. 144-173, June 2000.
12. Strzalkowski T., F. Lin, J. Perez-Carballo and J. Wang. Building Effective Queries in Natural language Information Retrieval. *Proceedings of the 5th Applied Natural Language Conference ANLP-97*, Washington D.C., USA, 1997.