Journal of Universal Computer Science, vol. 14, no. 11 (2008), 1877-1892
submitted: 30/9/07, accepted: 25/1/08, appeared: 1/6/08 © *J.UCS*

Structure and Semantics of Data-Intensive Web Pages: An Experimental Study on their Relationships

Lorenzo Blanco, Valter Crescenzi, Paolo Merialdo
(Università degli Studi Roma Tre, Roma, Italy
{blanco,crescenz,merialdo}@dia.uniroma3.it)

Abstract: In data-intensive web sites pages are generated by scripts that embed data from a back-end database into HTML templates. There is usually a relationship between the semantics of the data in a page and its corresponding template. For example, in a web site about sports events, it is likely that pages with data about athletes are associated with a template that differs from the template used to generate pages about coaches or referees. This article presents a method to classify web pages according to the associated template. Given a web page, the goal of our method is to accurately find the pages that are about the same topic. Our method leverages on a simple, yet effective model to abstract some structural features of a web page. We present the results of an extensive experimental analysis that show the performance of our methods in terms of both recall and precision regarding a large number of real-world web pages.
Key Words: Clustering, web page classification, data extraction
Category: H.3.3, I.2.6

1 Introduction

Nowadays, a large number of web sites provide information that is rendered by means of scripts that embed the results of a parametric database query into an HTML template. The semantics associated with the contents of a web page is usually related to the data returned by this query, which typically retrieves one instance of a conceptual entity from the database, e.g. an athlete or a stock quote; furthermore, the structure of a page is strictly related to the template from which it originates. As an example, consider the web pages about football players in Figure 1; the pages in Figure 2 come from the same web site, but they provide information about a football match, a football team, and some news. Although these pages have some common parts, e.g., headers and footers, the difference in their contents lead to substantial differences in the HTML presentation. It is reasonable to assume that player pages are generated by a template that is different from the template used to generate match or team pages. In general, since each HTML template is tailored to organise the results of a specific query, it is reasonable to expect that there exists a correlation between the structure of a page and its semantics.

This article investigates the effectiveness of template-based techniques for classifying web pages. We present a framework to model and compare the HTML structure of web pages, and prove that our techniques are effective enough by means of an extensive experimental study. Our proposal has many applications, namely: given a large wrapper library to extract and, possibly, integrate data from a large number of web sites, our techniques can be used to select the right wrapper in the library; furthermore, it can

Figure 1: Two web pages that publish data about football players.

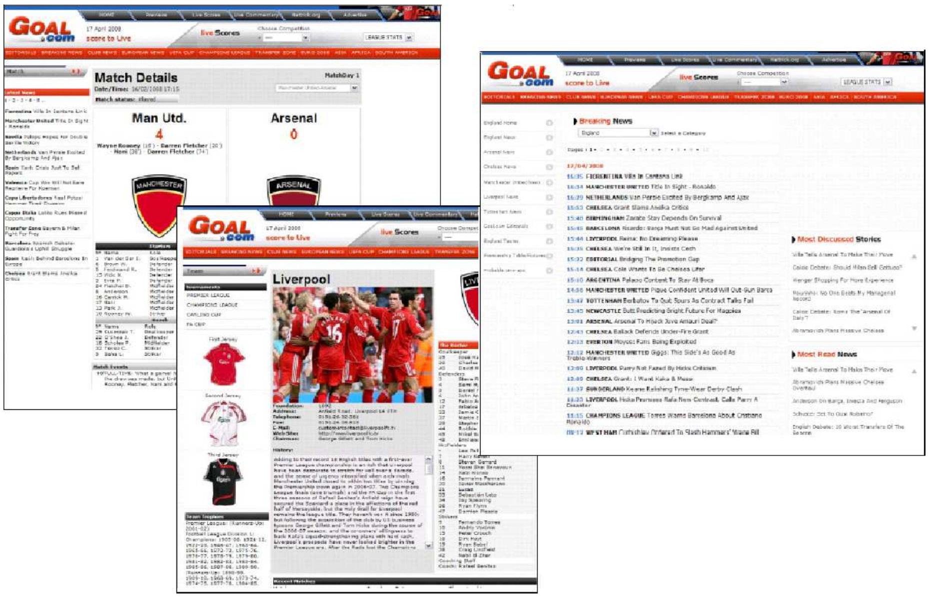

Figure 2: Web pages with data about football matches, teams, and sport news.

be used for text categorisation of pages from large data-intensive web sites. In general, given a small set of sample pages, our techniques can be used to determine which of pages in the same web site provide information about the same topic.

The rest of the article is organised as follows: we report on related proposals in Section 2; in Section 3, we define the problem and provide some definitions; in Section 4, we present the framework used to model and compare the HTML structure of web pages; in Section 5, we describe the experiments we have conducted to evaluate the framework; Section 6 concludes the article.

2 Related Work

The notion of page class to model sets of structurally homogeneous pages was introduced by [Atzeni et al., 2002] in the Araneus project, one of the pioneering approaches that proposed to model the logical structure of web sites for extraction purposes.

[Crescenzi et al., 2002], [Flesca et al., 2005], [Nierman and Jagadish, 2002], and also [Crescenzi et al., 2005] focus on clustering HTML web pages and XML documents according to their structure. [Crescenzi et al., 2002] and [Flesca et al., 2005] take a set of XML (possibly HTML) documents as input, and create clusters of pages based on a number of properties that are related to the frequency and the distribution of tags. [Flesca et al., 2005] investigated the use of similarity functions based on the Discrete Fourier Transform of documents encoded as time series, whereas the approach developed by [Nierman and Jagadish, 2002] is based on the more traditional concept of edit distance between trees that was also studied by [Zhang and Shasha, 1989]. The previous approaches measure the structural similarity between two XML documents, and are substantially different from ours, which measure the structural similarity between a document and a model that abstracts the template of a page class.

[Bertino et al., 2004] proposed an algorithm to measure the structural similarity between an XML document and a DTD; contrarily to our proposal, their definition of similarity is tailored to XML, not to HTML. Their algorithm can be used both to classify an XML document according to a library of given DTDs and to infer a representative DTD given a set of sample documents.

Our work is also related to the inference of a template from a set of positive samples. This issue has been studied in the context of information extraction [Turmo et al., 2006] and web data extraction [Laender et al., 2002] [Chang et al., 2006]. Relevant proposals in this field include [Crescenzi et al., 2001], [Crescenzi and Mecca, 2004], as well as [Arasu and Garcia-Molina, 2003], but the template models we use are substantially simpler, since we aim at classifying pages, not at extracting data from them. Other authors have proposed simpler template models, but their goals were different from ours; for instance [Ma et al., 2003] and [Bar-Yossef and Rajagopalan, 2002] focus on separating the contents of web pages from their template so as to improve the performances of information retrieval applications.

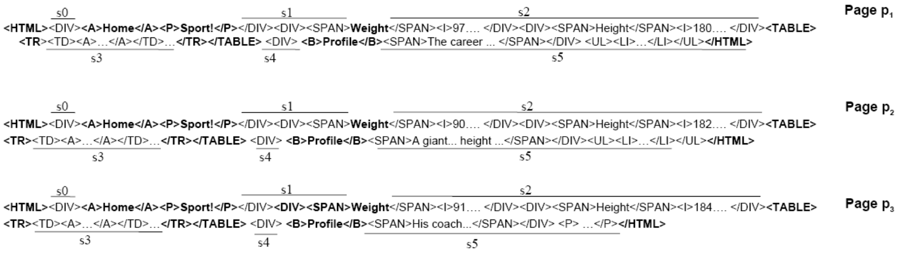

Figure 3: Sequences of tokens for three fictional web pages.

3 Problem Definition

Given a set of classes $\mathcal{C}$ and a set of documents $\mathcal{D}$, our goal is to determine which the class of each document $d \in \mathcal{D}$ is. We propose several classification schemes, each of which is associated with a page class model whose goal is to abstract a description of the pages in a given class $c \in \mathcal{C}$, and a similarity function, $sim(p, c)$, which returns a value in the range $[0, 1]$ that indicates how similar page p is to the pages in class c. The page class model for a given class c is computed from a set of sample pages that are classified by a human; such set is commonly referred to as the training set. A sample page is said to be positive for class c if it actually belongs to c, or negative otherwise. Some of the proposed page class models require the training set to include both positive and negative sample pages, whereas others work with positive samples only.

In our model, a web page is a sequence of tokens, which can be either HTML tags or strings. We take the name and the value of some HTML attributes into account, e.g., class and id; however, for the sake of simplicity, we refer to tags without any reference to their attributes. Each token t is associated with a pattern, denoted $pattern(t)$, which corresponds to the sequence of nodes from the root of the DOM tree of the page in which that token resides. A string token is associated with the pattern of the #PCDATA node in which it resides. Tokens t_1 and t_2 are *equivalent* if both $t_1 = t_2$, and $pattern(t_1) = pattern(t_2)$. Figure 3 depicts several web pages as sequences of tokens, and Figure 4 shows their corresponding DOM trees. In these pages, there are several equivalent tokens; for instance, tokens of the form $\langle \mathsf{I} \rangle$ are equivalent since they have the same pattern, i.e., HTML-DIV-I; furthermore, all of the occurrences of string token Weight are equivalent because their associated pattern is HTML-DIV-SPAN.

4 Classification Schemes

We propose four classification schemes, whose corresponding page class models have different expressiveness. The idea behind them is based on the observation that pages from data-intensive web sites are generated from a small number of HTML templates.

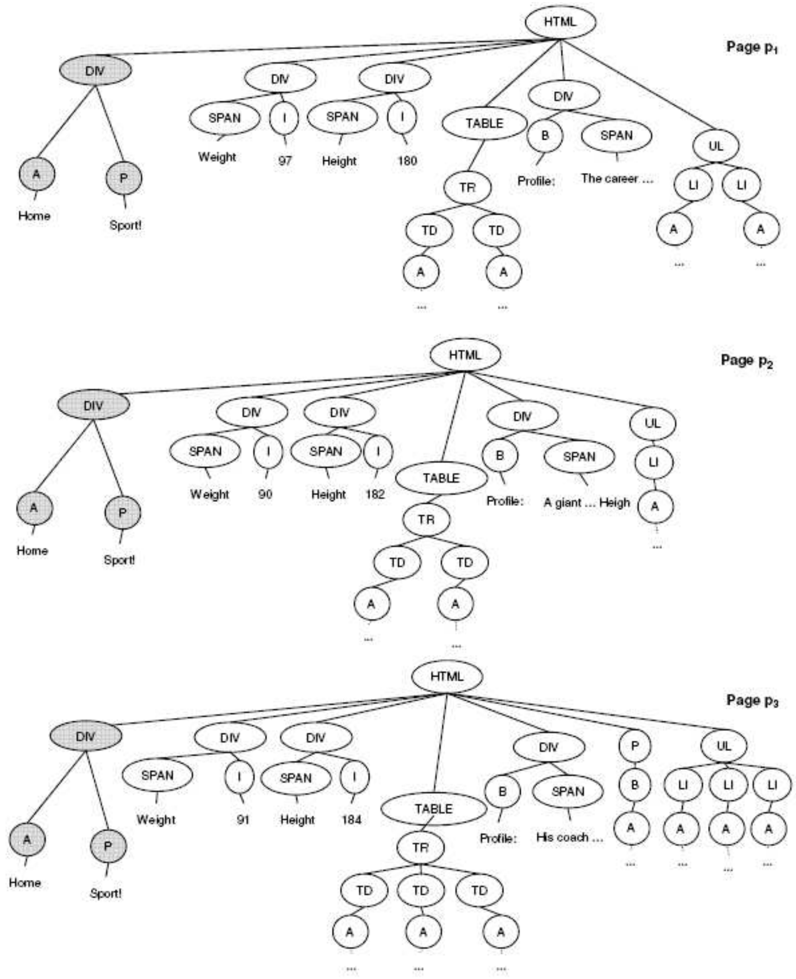

Figure 4: The DOM trees associated with the web pages in Figure 3.

Pages generated by the same template are structurally similar, whereas pages generated by different templates are clearly different from a structural point of view. The goal of our models is to describe the structure of a page by computing a subset of its DOM tree nodes that are strictly related to its underlying template.

The first classification scheme, which is called Template Page Model (TPM), uses a page class model that computes leaf nodes that are likely to belong to the same template. The second classification scheme, called Advanced Template Page Model (ATPM), improves on the previous scheme by excluding the nodes that belong to every page in a web site. The third classification scheme, called Link Page Model (LPM), uses a model

Algorithm 1 The TEMPLATETOKENS algorithm.

function TEMPLATETOKENS(S) ▷ $\mathcal{S} = \{s_1, \ldots, s_n\}$ is a set of sequences of tokens
  Let $\mathcal{T}$ be an empty set of tokens
  Let $\mathcal{E}_0 = \{e_1, \ldots, e_k\}$ be the list of tokens that occur exactly once in every element of $\mathcal{S}$
  for all token $e_i \in \mathcal{E}_0$ **do**
    Let $S^i = \{s^i_1, \ldots, s^i_n\}$ be a set of sequences
      such that $s^i_j = \text{SUBSEQUENCE}(s_j, \mathcal{E}, e_i)$ for all $j = 1, 2, \ldots, n$
    Add TEMPLATETOKENS(S^i) to $\mathcal{T}$
  end for
  return $\mathcal{T}$
end function

function SUBSEQUENCE($s, \mathcal{E}, e_i$) ▷ s a sequence of tokens $s = t_0 t_1 \ldots t_n$
  Let i be the index of e_i in s ▷ $\mathcal{E}$ a list of tokens $e_0, \ldots, e_k$; $e_j \in s, j = 1, \ldots, k$
  if $i = 0$ **then** ▷ e_i is a token, $e_i \in \mathcal{E}$
    $start \leftarrow 0$;
    $end \leftarrow i - 1$;
  end if
  if $i = k$ **then**
    $start \leftarrow i + 1$
    $end \leftarrow n$
  else
    $start \leftarrow i + 1$
    Let end be the index of e_{i+1} in s
    $end \leftarrow end - 1$
  end if
  return $t_{start} t_{start+1} \ldots t_{end}$
end function

that aims at describing the structure of a web page according the links it contains. LPM has an extension, called Advanced Link Page Model (ALPM), that excludes nodes that are common to every page in a web site. The LPM and the TPM models require positive sample pages only, whereas ATPM and ALPM need both positive and negative samples.

4.1 The Template Page Model Classification Scheme (TPM)

The TPM classification scheme introduces a page class model that aims at locating #PCDATA nodes that belong to a template by means of a statistical technique that builds on a proposal by [Arasu and Garcia-Molina, 2003]. Roughly speaking, given a set of pages generated by the same template, one can easily realise that the tokens in leaf nodes with the same pattern and frequency are likely to belong to the page template. For instance, consider the pages in Figure 3 again. Tokens such as Weight and Profile occur exactly once in these pages; it is reasonable then to assume that this is not by chance, but rather because they come from a template. Building on this idea, our goal is to seek for template tokens by selecting the set of nodes whose patterns occur once in every page of a given sample set. In the pages shown in Figure 4, the tokens that satisfy this condition are Home, Sport!, Weight and Profile:.

Note that the previous idea has a shortcoming, and that it must be refined to deal with it. For instance, token Height, which is likely to belong to the page template, cannot be included in the results because it occurs twice in the second page. To overcome this problem, we try to determine if the page can be split into homogeneous segments; in such cases, it is possible to inspect each segment recursively to find tokens that are not unique in the training set, but become unique within the context of a segment. For instance, notice that token Height occurs once in the third segment of every page, which is delimited by tokens Weight and ⟨TABLE⟩; therefore, it seems to be part of the template, whereas the other occurrence is incidentally equivalent.

Algorithm 1 provides a formal definition of procedure TEMPLATETOKENS, which computes the set of tokens that are likely to belong to a template. It extracts the tokens that are associated with patterns that occur once in each page of the training set; these tokens are then used to segment the input pages, and segments are recursively processed to discover other template tokens.

TPM Page Class Model

Given a set of pages P that are positive samples of class c, the page class model for c, denoted $\Delta^T(c)$, is the set of patterns associated with the tokens returned by procedure TEMPLATETOKENS(P). In our example, the TPM page class model computed from the pages in Figure 3 is as follows:

$$\begin{aligned}\Delta^T = \{&\mathsf{HTML\text{-}DIV\text{-}A\text{-}Home}, \mathsf{HTML\text{-}DIV\text{-}P\text{-}Sport!},\\ &\mathsf{HTML\text{-}DIV\text{-}SPAN\text{-}Weight}, \mathsf{HTML\text{-}DIV\text{-}SPAN\text{-}Height},\\ &\mathsf{HTML\text{-}DIV\text{-}B\text{-}Profile:}\}\end{aligned}$$

TPM Similarity Function

To measure how similar page p is to class c, we build on the fraction of patterns in $\Delta^T(c)$ that occur in p. In other words, if $X(p)$ denotes the set patterns associated with the tokens of p, then

$$sim^T(p,c) = \frac{|X(p) \cap \Delta^T(c)|}{|\Delta^T(c)|}$$

Notice that if p contains all of the patterns that characterise class c, then $sim^T(p,c)$ returns 1 since p clearly belongs to c; on the contrary, the similarity function returns 0 if p does not have any of the patterns in $\Delta^T(c)$.

4.2 The Advanced Template Page Model Classification Scheme (ATPM)

The Advanced Template Classification Scheme enhances the TPM scheme by discarding the patterns that are common to every page in a web site, e.g., headers, footers, or

Algorithm 2 The ADVANCEDTEMPLATETOKENS algorithm.

```
function ADVANCEDTEMPLATETOKENS(P, N)
    T ← TEMPLATETOKENS(P)          ▷ P is a set of positive sample pages
    S ← TEMPLATETOKENS(P ∪ N)      ▷ N is a set of negative sample pages
    return T − S
end function
```

navigational bars. To discard these patterns, we need the sample set to include some negative samples; usually, the home page is effective enough as a negative sample. By running TEMPLATETOKENS on a sample set that includes both positive and negative sample pages, we obtain patterns associated with tokens that are shared by a set of heterogeneous pages (typically by all the pages in that web site). These patterns are then excluded from the TPM schema computed on a positive sample set. Algorithm 2 defines this process formally.

ATPM Page Class Model

Given a set of pages P that includes both positive and negative samples for class c, the page class model for this class, denoted $\Delta^{T+}(c)$, is the set of patterns associated with the tokens returned by procedure ADVANCEDTEMPLATETOKENS(P). In our example, assuming that the gray nodes in Figure 4 represent nodes that also belong to the negative sample pages, the ATPM page class model is as follows:

$$\Delta^{T+}(c) = \{\text{HTML-DIV-SPAN-Weight}, \text{HTML-DIV-SPAN-Height}, \text{HTML-DIV-B-Profile:}\}$$

ATPM Similarity Function

The similarity function for the ATPM model is defined as the fraction of patterns in $\Delta^{T+}(c)$ that also occur in p. In other words, if $X(p)$ denotes the set patterns associated with the tokens of p, then:

$$sim^{T+}(p,c) = \frac{|X(p) \cap \Delta^{T+}(c)|}{|\Delta^{T+}(c)|}$$

4.3 The Link Page Model Classification Scheme (LPM)

The Link Page Model classification scheme is based on a page class model that deviates from the previous ones since we now focus on the links in the pages in the training set. This model is motivated by the following observations: (i) pages from large data-intensive web sites usually contain a large number of links; (ii) the set of layout and presentation properties associated with the links of a page are related to the structure of the page itself since different templates produce different links. If the majority of links in two pages share the same layout and presentation properties, it is then likely that these two pages share the same structure.

LPM Page Class Model

Given a set of pages P that has positive samples of class c only, its page class model, denoted $\Delta^L(c)$, is the set of patterns associated with tokens of the form $\langle\mathsf{A}\rangle$ that occur at least once in every page of P. For instance, regarding the pages in Figure 4, the LPM page class model is as follows:

$$\Delta^L(c) = \{\text{HTML-DIV-A}, \text{HTML-TABLE-TR-TD-A}, \\ \text{HTML-UL-LI-A}\}$$

It is worth observing that, differently from ATPM and TPM, the LPM page class model can be computed even with a singleton sample set. For example, the LPM page class model for the singleton set $P = \{p_3\}$ is as follows:

$$\Delta^L(c) = \{\text{HTML-DIV-A}, \text{HTML-TABLE-TR-TD-A}, \\ \text{HTML-UL-LI-A}, \text{HTML-P-B-A}\}$$

LPM Similarity Function

The similarity between page p and class c is computed by comparing the set of patterns in $\Delta^L(c)$ and the set of patterns associated to the links in p. We use the Jaccard coefficient to compare them [Hand et al., 2001]. In other words, if $X(p)$ denotes the set patterns associated with the $\langle\mathsf{A}\rangle$ tokens of p, then

$$sim^L(p,c) = \frac{|X(p) \cap \Delta^L(c)|}{|X(p) \cup \Delta^L(c)|}$$

4.4 The Advanced Link Page Model Classification Scheme (ALPM)

The Advanced Link Page Model classification scheme is a variant of LPM in which $\langle\mathsf{A}\rangle$ nodes that belong to every page in a web site are discarded.

ALPM Page Class Model

Given a set of pages P that includes both positive and negative samples of class c, its page class model, denoted $\Delta^{L^+}(c)$, is the set of patterns associated with tokens of the form $\langle\mathsf{A}\rangle$ that occur at least once in every positive sample page of P, minus the set of such patterns that occur at least once in at least a negative sample page of P. In our example, assuming that the gray nodes in Figure 4 represent nodes that also belong to the negative sample pages, the ALPM page class model is as follows:

$$\Delta^L(c) = \{\text{HTML-TABLE-TR-TD-A}, \text{HTML-UL-LI-A}\}$$

Classification Scheme	Number of Positive Samples	Number of Negative Samples
TPM	4	0
ATPM	4	1
LPM	1	0
ALPM	1	1

Table 1: Features of training sets.

ALPM Similarity Function

This function is very similar to the previous case. If $X(p)$ denotes the set patterns associated with the $\langle \mathsf{A} \rangle$ tokens of p, then

$$sim^{L+}(p,c) = \frac{|X(p) \cap \Delta^{L+}(c)|}{|X(p) \cup \Delta^{L+}(c)|}$$

5 Experiments

In this section, we report on the experiments we conducted to evaluate the effectiveness of our classification schemes. The experiments were run on real-world pages from 125 data-intensive web sites. For each site, we identified one class of pages and gathered several pages of that class. For each class, we built four training sets, one for each classification scheme. Table 1 shows the number of samples in the training sets. The positive examples were chosen by picking a small number of samples at random, whereas the negative example was always the home page of the corresponding web site. In the case of the LPM classification scheme, the training set was composed of just one positive example. For each class, we then built a test set that was composed of pages that were labelled by a person. Our training sets were composed of about 25 positive samples and 15 negative samples, and they were disjoint from the test sets. For each page in the test set we have computed the values returned by the similarity function associated to each classification scheme: if the similarity was greater than a threshold value, which was empirically set to 0.75, the page was then considered to be a member of the target class.

We also built two baseline classifiers using Naive Bayes [Manning et al., 2008], which were trained with pages in the training set used for the ATPM classification scheme. The former classifier modelled web pages as sequences of tokens according to the well-known bag of words model [Manning et al., 2008]; the later, relied on both string tokens and patterns so that terms with different patterns in the DOM tree were not considered equivalent, i.e., the role of the tokens in the HTML was taken into account.

	Naive Bayes			Naive Bayes+			LPM			ALPM			TPM			ATPM		
Domain	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**
Finance	0.59	0.97	0.71	0.75	0.86	0.78	0.98	0.87	0.90	0.99	0.73	0.79	0.94	0.98	0.95	1.00	0.98	0.99
Films	0.55	0.94	0.64	0.71	0.81	0.72	0.97	0.78	0.82	0.98	0.65	0.71	0.86	0.97	0.89	0.94	0.97	0.94
Football	0.61	0.98	0.72	0.79	0.84	0.77	1.00	0.86	0.89	0.99	0.76	0.81	0.92	0.95	0.92	0.98	0.95	0.95
Basketball	0.53	1.00	0.67	0.74	0.81	0.73	0.94	0.92	0.90	0.99	0.82	0.86	0.83	0.99	0.89	0.94	0.98	0.96
Average	**0.57**	**0.97**	**0.69**	**0.75**	**0.83**	**0.75**	**0.97**	**0.86**	**0.88**	**0.99**	**0.74**	**0.79**	**0.89**	**0.97**	**0.91**	**0.97**	**0.97**	**0.96**

Table 2: Summary of the classification results.

5.1 Evaluation Metrics

The performance of our schemes was evaluated by means of the standard precision, recall, and F-measure metrics. Precision measures the ratio of pages classified as positive that are actually positive pages, i.e., $p = \frac{tp}{tp+fp}$, where *tp* is the number of pages that are correctly classified as positive and *fp* is the number of negative examples that are incorrectly classified as positive; recall measures the ratio of pages that are correctly labelled, i.e., $r = \frac{tp}{tp+fn}$, where *fn* refers to the number of pages incorrectly labelled as negative (note that $tp+fn$ is the total number of examples of the test set); the F-measure is defined as the harmonic mean of precision and recall, i.e., $F = 2\frac{p \cdot r}{p+r}$; intuitively, the higher the F-measure, the better the performance.

5.2 Results

Table 2 summarises our results on a set of 5 000 pages from the following domains: finance, films, football, and basketball. The details are presented in Tables 3, 4, 5, and 6, respectively. Notice that our classification schemes outperform both Naive Bayes classifiers, and that the one that takes patterns into account performs better that the other, which confirms that taking structural features into account improves classification.

The ATPM classification scheme clearly outperforms the others in almost all cases. The value of the F-measure for the LPM scheme is higher than for the ALPM scheme. Looking at the individual contributions of precision and recall, we realised that the ALPM scheme is more precise (even more than ATPM), but its recall is significantly lower. In this classification scheme, the presence or absence of a few links has a heavy impact since, for classes with a small number of characteristic links, the classifier discards many pages, even if their links differ very little from the page model.

Although our results are quite good in general, there are a few cases in which the Naive Bayes classifiers perform better than our schemes. For example, our schemes do not perform well in site http://www.uit.no, cf. Table 5, which is a university site with a section that is devoted to some football players; after inspecting these pages, we realised that they were organised according to very simple templates, and, therefore, the impact of structure was negligible with respect to the contents. The classification

	Naive Bayes			Naive Bayes+			LPM			ALPM			TPM			ATPM		
Web Site	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**
finance.google.com	0.15	1.00	0.26	0.75	1.00	0.86	1.00	1.00	1.00	1.00	0.33	0.50	1.00	1.00	1.00	1.00	1.00	1.00
finance.yahoo.com	0.21	1.00	0.35	0.43	1.00	0.60	0.91	0.96	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
marketwatch.com	0.25	1.00	0.40	0.83	1.00	0.91	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
money.cnn.com	0.47	1.00	0.64	0.89	1.00	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
markets.about.com	0.50	0.80	0.62	0.52	0.75	0.61	1.00	0.10	0.18	1.00	0.10	0.18	1.00	1.00	1.00	1.00	1.00	1.00
studio-5.financialcontent.com	0.51	0.90	0.65	0.60	0.75	0.67	1.00	0.65	0.79	1.00	0.25	0.40	1.00	1.00	1.00	1.00	1.00	1.00
finance.abc7news.com	0.53	0.95	0.68	0.70	0.80	0.74	1.00	0.93	0.96	1.00	0.15	0.26	1.00	1.00	1.00	1.00	1.00	1.00
stocks.us.reuters.com	0.53	1.00	0.69	0.81	0.85	0.83	1.00	0.95	0.97	1.00	0.78	0.88	1.00	0.95	0.97	1.00	0.95	0.97
finance.banks.com	0.53	0.95	0.68	0.57	0.89	0.69	1.00	0.95	0.97	1.00	0.50	0.67	1.00	1.00	1.00	1.00	1.00	1.00
smartmoney.com	0.54	1.00	0.70	0.80	1.00	0.89	1.00	0.95	0.97	1.00	0.95	0.97	1.00	1.00	1.00	1.00	1.00	1.00
finance.sfgate.com	0.55	0.90	0.68	0.61	0.85	0.71	1.00	0.80	0.89	1.00	0.80	0.89	0.57	1.00	0.73	1.00	1.00	1.00
fin-rus.com	0.57	1.00	0.72	0.60	1.00	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
caps.fool.com	0.63	1.00	0.77	0.67	1.00	0.80	1.00	0.95	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
beta.finance.aol.com	0.63	1.00	0.77	0.94	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
quote.barchart.com	0.63	1.00	0.78	0.95	1.00	0.98	0.87	1.00	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
finance.dmwmedia.com	0.66	0.95	0.78	0.73	0.95	0.83	1.00	0.90	0.95	1.00	0.10	0.18	0.67	1.00	0.80	1.00	1.00	1.00
otcjournal.com	0.66	1.00	0.79	0.71	1.00	0.83	0.71	1.00	0.83	0.68	0.90	0.78	0.64	1.00	0.78	1.00	0.97	0.98
techweb.com	0.67	1.00	0.80	0.50	0.25	0.33	1.00	0.75	0.86	1.00	0.90	0.95	1.00	1.00	1.00	1.00	1.00	1.00
finance.optimum.net	0.70	0.89	0.78	0.85	0.85	0.85	1.00	0.78	0.88	1.00	0.78	0.88	1.00	1.00	1.00	1.00	1.00	1.00
quote.morningstar.com	0.73	1.00	0.84	0.85	0.96	0.90	1.00	0.92	0.96	1.00	0.96	0.98	1.00	0.96	0.98	1.00	0.96	0.98
finance.paidcontent.org	0.76	0.95	0.84	0.83	0.50	0.62	1.00	0.80	0.89	1.00	0.80	0.89	0.77	1.00	0.87	1.00	0.95	0.97
investing.businessweek.com	0.83	1.00	0.91	0.88	0.92	0.90	1.00	0.92	0.96	1.00	0.77	0.87	0.84	1.00	0.91	1.00	1.00	1.00
amex.com	0.89	0.98	0.93	0.95	0.50	0.66	1.00	0.57	0.73	1.00	0.60	0.75	1.00	0.60	0.75	1.00	0.60	0.75
uk.finance.yahoo.com	0.96	1.00	0.98	0.89	0.62	0.73	1.00	0.96	0.98	1.00	0.95	0.97	1.00	1.00	1.00	1.00	1.00	1.00
investorguide.com	0.67	1.00	0.80	0.93	1.00	0.96	1.00	0.80	0.89	1.00	0.69	0.82	1.00	1.00	1.00	1.00	1.00	1.00
Average	**0.58**	**0.97**	**0.69**	**0.76**	**0.83**	**0.76**	**0.98**	**0.81**	**0.85**	**0.99**	**0.70**	**0.76**	**0.90**	**0.97**	**0.92**	**0.98**	**0.97**	**0.97**

Table 3: Results for pages in the finance domain (target pages: stock quotes).

schemes based on links lead to low values for both precision and recall in cases such as http://www.eccentric-cinema.com, cf. Table 4, whereas the TPM and ATPM classification schemes perform very well; in this case, the problem is due to the fact that the pages in this site have very few links.

Overall the LPM scheme can be considered a good trade-off between efficiency and effectiveness. It produces precise results, with good recall; furthermore, compared to the other schemes, it requires a training set composed of just one positive example.

6 Conclusions

We have proposed several classification schemes, and we have compared their performance with two Naive Bayes classifiers. Our methods have produced interesting results, both in terms of precision and recall. We have also realised that the structure of a page is actually related to its semantics in data-intensive web sites. That is, it is possible to produce a semantic classification of pages by analysing their structural features. Our

	Naive Bayes			**Naive Bayes+**			**LPM**			**ALPM**			**TPM**			**ATPM**		
Web Site	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**
movies.aol.com	0.21	1.00	0.34	0.14	0.43	0.21	1.00	0.86	0.92	1.00	0.70	0.82	0.17	0.86	0.29	1.00	0.86	0.92
movies.go.com	0.58	1.00	0.73	0.92	0.55	0.69	1.00	0.60	0.75	1.00	0.15	0.26	1.00	0.75	0.86	1.00	0.55	0.71
movies.msn.com	0.57	1.00	0.72	0.74	1.00	0.85	1.00	0.69	0.82	1.00	0.48	0.65	1.00	1.00	1.00	0.88	0.9388	0.91
movies.nytimes.com	0.11	1.00	0.20	0.50	1.00	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
movies.yahoo.com	0.58	1.00	0.74	0.61	1.00	0.76	1.00	0.84	0.91	1.00	0.76	0.86	1.00	0.96	0.98	1.00	0.96	0.98
rogerebert.suntimes.com	0.68	1.00	0.81	0.50	0.15	0.24	1.00	1.00	1.00	1.00	0.65	0.79	0.81	1.00	0.90	1.00	1.00	1.00
uk.movies.yahoo.com	0.29	1.00	0.44	0.38	0.50	0.43	1.00	0.33	0.50	1.00	0.33	0.50	0.75	1.00	0.86	0.75	1.00	0.86
all-reviews.com	0.07	1.00	0.13	0.25	1.00	0.40	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
bigscreen.com	0.60	1.00	0.75	0.66	0.81	0.72	1.00	1.00	1.00	1.00	0.08	0.14	1.00	1.00	1.00	1.00	1.00	1.00
boxofficemojo.com	0.58	1.00	0.73	0.74	1.00	0.85	1.00	0.08	0.14	1.00	0.08	0.14	1.00	1.00	1.00	1.00	1.00	1.00
cinema.com	0.67	1.00	0.80	0.74	1.00	0.85	0.46	1.00	0.63	0.46	1.00	0.63	0.46	1.00	0.63	0.46	1.00	0.63
cinemareview.com	0.78	1.00	0.88	0.95	0.72	0.82	1.00	0.79	0.88	1.00	0.56	0.72	1.00	0.68	0.81	1.00	0.68	0.81
comingsoon.net	0.61	1.00	0.76	0.90	0.73	0.81	1.00	0.96	0.98	1.00	0.81	0.89	1.00	1.00	1.00	1.00	1.00	1.00
dtmdb.com	0.25	1.00	0.40	0.18	0.67	0.29	1.00	1.00	1.00	1.00	1.00	1.00	0.24	1.00	0.39	0.39	1.00	0.56
eccentric-cinema.com	0.94	0.58	0.71	0.73	0.67	0.70	1.00	0.04	0.07	1.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
fandango.com	0.16	1.00	0.27	0.50	1.00	0.67	1.00	1.00	1.00	1.00	1.00	1.00	0.60	1.00	0.75	1.00	1.00	1.00
filmcritic.com	0.85	1.00	0.92	0.92	0.94	0.93	1.00	0.91	0.96	1.00	0.80	0.89	0.83	1.00	0.91	0.85	1.00	0.92
filmsite.org	0.38	0.12	0.18	0.94	0.58	0.71	1.00	0.46	0.63	1.00	0.73	0.84	1.00	1.00	1.00	1.00	1.00	1.00
imdb.com	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
listentoamovie.com	0.53	1.00	0.69	0.63	1.00	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
metacritic.com	0.72	1.00	0.84	0.95	0.81	0.87	1.00	0.96	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
movietome.com	0.81	1.00	0.90	0.74	0.54	0.62	0.80	0.15	0.26	0.82	0.05	0.09	0.96	1.00	0.98	0.96	1.00	0.98
movieweb.com	0.68	1.00	0.81	0.93	0.96	0.94	1.00	0.81	0.89	1.00	0.42	0.59	1.00	1.00	1.00	1.00	1.00	1.00
reelviews.net	0.94	0.58	0.71	0.93	0.56	0.70	1.00	0.65	0.79	1.00	0.55	0.71	1.00	1.00	1.00	1.00	1.00	1.00
rottentomatoes.com	0.67	1.00	0.80	0.67	0.77	0.71	1.00	1.00	1.00	1.00	0.81	0.89	1.00	1.00	1.00	1.00	1.00	1.00
starpulse.com	0.63	1.00	0.78	0.93	1.00	0.96	1.00	0.96	0.98	1.00	0.69	0.82	1.00	1.00	1.00	1.00	1.00	1.00
time.com	0.37	1.00	0.54	0.83	1.00	0.91	1.00	0.91	0.95	1.00	0.91	0.95	1.00	1.00	1.00	1.00	1.00	1.00
tvguide.com	0.26	1.00	0.41	0.75	1.00	0.86	1.00	0.50	0.67	1.00	0.17	0.29	0.43	1.00	0.60	1.00	1.00	1.00
webwombat.com.au	0.36	1.00	0.53	0.89	1.00	0.94	1.00	1.00	1.00	1.00	1.00	1.00	0.73	1.00	0.84	0.89	1.00	0.94
Average	**0.58**	**0.97**	**0.68**	**0.76**	**0.82**	**0.75**	**0.97**	**0.83**	**0.85**	**0.99**	**0.72**	**0.77**	**0.89**	**0.96**	**0.90**	**0.97**	**0.96**	**0.95**

Table 4: Results for pages in the film domain (target pages: films).

methods can then be used to support web wrapper systems, for example to determine which the best wrapper to deal with a given page is, or to help crawlers perform better, which is currently being explored in the Flint project [Blanco et al., 2008].

	Naive Bayes			Naive Bayes+			LPM			ALPM			TPM			ATPM		
Web Site	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**
arsenal-mania.com	0.48	1.00	0.65	0.88	1.00	0.93	1.00	1.00	1.00	1.00	0.93	0.96	1.00	1.00	1.00	1.00	1.00	1.00
chivas.usa.mlsnet.com	0.73	1.00	0.84	1.00	0.50	0.67	1.00	0.25	0.40	1.00	0.10	0.18	0.80	1.00	0.89	1.00	1.00	1.00
columbus.crew.mlsnet.com	0.65	1.00	0.79	0.91	0.77	0.83	1.00	0.77	0.87	1.00	0.23	0.38	1.00	0.85	0.92	1.00	0.77	0.87
dcunited.mlsnet.com	0.43	1.00	0.60	0.83	0.77	0.80	0.91	0.77	0.83	0.96	0.67	0.79	1.00	1.00	1.00	1.00	1.00	1.00
expertfootball.com	0.44	0.66	0.53	0.86	0.38	0.52	1.00	0.33	0.50	0.66	0.66	0.66	0.71	0.83	0.77	0.70	0.83	0.76
fifa02	0.07	1.00	0.13	0.50	1.00	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
football.stats.com	1.00	0.92	0.96	1.00	0.92	0.96	1.00	0.92	0.96	1.00	0.92	0.96	1.00	1.00	1.00	1.00	1.00	1.00
footballdatabase.com	0.48	1.00	0.65	1.00	0.80	0.89	1.00	0.07	0.13	1.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
msn.foxsports.com	0.10	1.00	0.18	0.33	1.00	0.50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
redbull.newyork.mlsnet.com	0.94	1.00	0.97	1.00	0.88	0.93	1.00	0.69	0.81	1.00	0.69	0.82	1.00	0.94	0.97	1.00	0.94	0.97
soccer.azplayers.com	0.99	0.94	0.96	0.94	1.00	0.97	1.00	1.00	1.00	1.00	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00
soccernet-akamai.espn.go.com	0.46	1.00	0.63	0.81	1.00	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
soccernet.espn.go.com	0.51	1.00	0.68	0.95	1.00	0.97	1.00	0.42	0.59	1.00	0.42	0.59	1.00	1.00	1.00	1.00	1.00	1.00
soccertimes.com	0.57	1.00	0.73	0.92	0.92	0.92	1.00	1.00	1.00	1.00	0.83	0.91	1.00	0.92	0.96	1.00	0.92	0.96
uk.eurosport.yahoo.com	0.06	1.00	0.11	0.50	1.00	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ussoccerplayers.com	0.95	1.00	0.98	1.00	1.00	1.00	1.00	0.29	0.44	1.00	0.20	0.33	1.00	1.00	1.00	1.00	1.00	1.00
worldsoccer.about.com	0.81	1.00	0.89	0.95	0.86	0.90	1.00	0.71	0.83	1.00	0.71	0.83	1.00	1.00	1.00	1.00	1.00	1.00
4thegame.com	0.55	1.00	0.71	0.92	0.65	0.76	1.00	0.24	0.38	1.00	0.10	0.18	0.81	1.00	0.89	1.00	1.00	1.00
acmilan.com	0.10	1.00	0.18	0.10	1.00	0.18	1.00	1.00	1.00	1.00	1.00	1.00	0.50	1.00	0.67	1.00	1.00	1.00
carling.com	0.55	1.00	0.71	0.75	1.00	0.86	1.00	1.00	1.00	1.00	0.83	0.91	1.00	1.00	1.00	1.00	1.00	1.00
englandfootballonline.com	0.75	1.00	0.86	0.20	0.07	0.10	1.00	1.00	1.00	1.00	0.78	0.88	0.67	0.67	0.67	1.00	1.00	1.00
evertonfc.com	0.56	1.00	0.72	0.68	1.00	0.81	1.00	1.00	1.00	1.00	1.00	1.00	0.58	1.00	0.73	1.00	1.00	1.00
fcbayern.t-com.de	0.77	1.00	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
football-rumours.com	0.68	1.00	0.81	0.70	1.00	0.82	1.00	0.95	0.98	1.00	0.86	0.92	0.51	1.00	0.68	1.00	1.00	1.00
football.co.uk	0.60	1.00	0.75	1.00	0.76	0.86	1.00	0.95	0.98	1.00	0.95	0.98	1.00	1.00	1.00	1.00	1.00	1.00
footballplus.com	0.78	1.00	0.88	0.95	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
fulhamfc.com	0.74	1.00	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
galatasaray.org	0.47	1.00	0.64	0.76	1.00	0.86	1.00	1.00	1.00	1.00	1.00	1.00	0.46	1.00	0.63	1.00	0.94	0.97
gfdb.com	0.81	1.00	0.89	0.94	0.81	0.87	1.00	0.95	0.98	1.00	0.95	0.98	1.00	1.00	1.00	1.00	1.00	1.00
goal.com	0.52	1.00	0.68	1.00	1.00	1.00	1.00	0.93	0.97	1.00	0.70	0.82	1.00	1.00	1.00	1.00	1.00	1.00
juventus.com	0.50	1.00	0.67	0.50	0.08	0.14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
leedsfans.org.uk	0.83	1.00	0.90	1.00	0.47	0.64	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
liverpoolfc.tv	0.95	0.98	0.97	0.98	0.68	0.80	1.00	1.00	1.00	1.00	0.42	0.59	1.00	0.68	0.81	1.00	0.68	0.81
newcastle-online.com	0.65	0.75	0.70	1.00	0.75	0.86	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.75	0.86	1.00	0.75	0.86
nufc.premiumtv.co.uk	0.68	1.00	0.81	0.23	0.14	0.18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
officialplayersites.com	0.62	1.00	0.76	1.00	0.67	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
paktribune.com	0.08	1.00	0.15	0.10	1.00	0.18	1.00	1.00	1.00	1.00	1.00	1.00	0.50	1.00	0.67	0.50	1.00	0.67
proplayermanagement.com	0.72	1.00	0.84	0.98	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
soccer-gallery.net	0.39	1.00	0.56	0.41	1.00	0.58	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
soccerbase.com	0.88	1.00	0.94	0.75	1.00	0.86	1.00	1.00	1.00	1.00	0.20	0.33	1.00	1.00	1.00	1.00	1.00	1.00
uit.no	0.85	1.00	0.92	0.83	0.91	0.87	1.00	0.18	0.31	1.00	0.00	0.00	1.00	0.09	0.17	1.00	0.09	0.17
us.terra.com	0.65	1.00	0.79	0.76	1.00	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
werder.de	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.56	0.71	1.00	1.00	1.00	1.00	1.00	1.00
wldcup.com	0.66	1.00	0.79	0.95	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
yanks-abroad.com	0.88	1.00	0.94	0.85	0.89	0.87	1.00	1.00	1.00	1.00	0.33	0.49	1.00	0.99	0.99	1.00	0.99	0.99
zerozero.pt	0.37	1.00	0.54	0.65	1.00	0.79	1.00	1.00	1.00	1.00	0.91	0.95	0.73	1.00	0.85	1.00	0.82	0.90
Average	**0.59**	**0.99**	**0.70**	**0.77**	**0.83**	**0.75**	**0.97**	**0.88**	**0.88**	**0.99**	**0.76**	**0.80**	**0.89**	**0.96**	**0.90**	**0.97**	**0.95**	**0.94**

Table 5: Results for pages in the football domain (target pages: football players).

	Naive Bayes			Naive Bayes+			LPM			ALPM			TPM			ATPM		
Web Site	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**	**p**	**r**	**F**
cbs.sportsline.com	0.63	1.00	0.77	0.84	0.73	0.78	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
fantasybasketball.usatoday.com	0.80	1.00	0.89	0.89	1.00	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
fightingillini.cstv.com	0.45	1.00	0.62	0.67	0.92	0.77	1.00	1.00	1.00	1.00	1.00	1.00	0.50	1.00	0.67	1.00	1.00	1.00
gozags.cstv.com	0.33	1.00	0.50	1.00	0.88	0.93	1.00	1.00	1.00	1.00	0.88	0.93	1.00	1.00	1.00	1.00	1.00	1.00
msuspartans.cstv.com	0.34	1.00	0.51	0.90	0.90	0.90	1.00	1.00	1.00	1.00	1.00	1.00	0.91	1.00	0.95	1.00	1.00	1.00
robots.cnnsi.com	0.36	1.00	0.53	1.00	0.80	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
scoreboards.aol.com	0.23	1.00	0.37	1.00	0.13	0.22	0.28	1.00	0.44	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
sports.espn.go.com	0.34	1.00	0.51	0.50	0.70	0.58	1.00	0.40	0.57	1.00	0.24	0.39	0.43	0.90	0.58	0.41	0.70	0.52
stats.globesports.com	0.42	1.00	0.59	0.45	1.00	0.62	0.47	1.00	0.64	1.00	1.00	1.00	0.42	1.00	0.59	1.00	1.00	1.00
uclabruins.cstv.com	0.39	1.00	0.56	0.44	1.00	0.61	1.00	1.00	1.00	1.00	1.00	1.00	0.85	1.00	0.92	1.00	1.00	1.00
und.cstv.com	0.39	1.00	0.56	0.73	0.89	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
usabasketball.com	0.66	1.00	0.79	0.77	0.65	0.71	1.00	1.00	1.00	1.00	0.69	0.82	0.81	1.00	0.90	1.00	1.00	1.00
wakeforestsports.cstv.com	0.36	1.00	0.53	0.28	0.50	0.36	1.00	1.00	1.00	1.00	0.30	0.46	0.50	1.00	0.67	1.00	1.00	1.00
basketball-reference.com	0.93	1.00	0.96	0.60	1.00	0.75	0.71	0.96	0.82	0.72	1.00	0.84	0.60	1.00	0.75	0.60	1.00	0.75
collegehoopsnet.com	0.58	1.00	0.74	1.00	0.29	0.44	1.00	0.71	0.83	1.00	0.53	0.69	0.70	1.00	0.82	0.72	1.00	0.84
covers.com	0.70	1.00	0.83	0.72	1.00	0.84	1.00	0.31	0.47	1.00	0.10	0.18	0.65	1.00	0.79	0.93	1.00	0.96
databasebasketball.com	0.68	1.00	0.81	0.57	0.65	0.61	1.00	0.81	0.89	1.00	0.72	0.84	1.00	0.88	0.94	1.00	0.88	0.94
hoopshype.com	0.80	1.00	0.89	0.81	0.88	0.84	1.00	0.92	0.96	1.00	0.77	0.87	0.73	1.00	0.84	1.00	1.00	1.00
hoopsstats.com	0.33	1.00	0.50	0.58	1.00	0.73	1.00	1.00	1.00	1.00	1.00	1.00	0.92	1.00	0.96	1.00	1.00	1.00
hoopsvibe.com	0.72	1.00	0.84	0.80	0.77	0.78	1.00	1.00	1.00	1.00	0.77	0.87	1.00	1.00	1.00	1.00	1.00	1.00
kfba.net	0.64	1.00	0.78	0.86	0.86	0.86	1.00	1.00	1.00	1.00	0.57	0.73	1.00	1.00	1.00	1.00	1.00	1.00
rototimes.com	0.43	1.00	0.60	0.90	1.00	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
rotoworld.com	0.58	1.00	0.73	0.87	1.00	0.93	1.00	1.00	1.00	1.00	1.00	1.00	0.90	1.00	0.95	0.90	1.00	0.95
www2.sportsnet.ca	0.62	1.00	0.77	0.65	0.94	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Average	**0.52**	**1.00**	**0.67**	**0.72**	**0.81**	**0.72**	**0.91**	**0.92**	**0.88**	**0.96**	**0.78**	**0.82**	**0.84**	**0.96**	**0.86**	**0.94**	**0.94**	**0.92**

Table 6: Results for pages in the basketball domain (target pages: basketball players).

References

[Arasu and Garcia-Molina, 2003] Arasu, A. and Garcia-Molina, H. (2003). Extracting structured data from web pages. In *Proceedings of the 19th Conference on Management of Data*, pages 337–348.

[Atzeni et al., 2002] Atzeni, P., Mecca, G., and Merialdo, P. (2002). Managing web-based data: Database models and transformation. *IEEE Internet Computing*, 6(4):33–37.

[Bar-Yossef and Rajagopalan, 2002] Bar-Yossef, Z. and Rajagopalan, S. (2002). Template detection via data mining and its application. In *Proceedings of the 11th World Wide Web Conference*, pages 580–591.

[Bertino et al., 2004] Bertino, E., Guerrini, G., and Mesiti, M. (2004). A matching algorithm for measuring the structural similarity between an XML document and a DTD and its application. *Information Systems*, 29(1):23–46.

[Blanco et al., 2008] Blanco, L., Crescenzi, V., Merialdo, P., and Papotti, P. (2008). Flint: Google-basing the Web. In *Proceedings of the 16th Conference on Extending Database Technology*, pages 720–724.

[Chang et al., 2006] Chang, C.-H., Kayed, M., Girgis, M., and Shaalan, K. (2006). A survey of web information extraction systems. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1411–1428.

[Crescenzi and Mecca, 2004] Crescenzi, V. and Mecca, G. (2004). Automatic information extraction from large web sites. *Journal of the ACM*, 51(5):731–779.

[Crescenzi et al., 2001] Crescenzi, V., Mecca, G., and Merialdo, P. (2001). ROADRUNNER: Towards automatic data extraction from large web sites. In *Proceedings of the 20th Conference on Very Large Data Bases*, pages 109–118.

[Crescenzi et al., 2002] Crescenzi, V., Mecca, G., and Merialdo, P. (2002). Wrapping-oriented classification of web pages. In *Proceedings of the 2002 ACM Symposium on Applied Computing*, pages 1108–1112.

[Crescenzi et al., 2005] Crescenzi, V., Merialdo, P., and Missier, P. (2005). Clustering pages based on their structure. *Data and Knowledge Engineering*, 54(3):279–299.

[Flesca et al., 2005] Flesca, S., Manco, G., Masciari, E., Pontieri, L., and Pugliese, A. (2005). Fast detection of XML structural similarity. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):160–175.

[Hand et al., 2001] Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. The MIT Press.

[Laender et al., 2002] Laender, A., Ribeiro-Neto, B., da Silva, A., and Teixeira, J. (2002). A brief survey of web data extraction tools. *ACM SIGMOD Record*, 31(2):84–93.

[Ma et al., 2003] Ma, L., Goharian, N., Chowdhury, A., and Chung, M. (2003). Extracting unstructured data from template generated web documents. In *Proceedings of the 12th Conference on Information and Knowledge Management*, pages 512–515.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

[Nierman and Jagadish, 2002] Nierman, A. and Jagadish, H. (2002). Evaluating structural similarity in XML document. In *Proceedings of the 5th Workshop on the Web and Databases*, pages 61–66.

[Turmo et al., 2006] Turmo, J., Ageno, A., and Català, N. (2006). Adaptive information extraction. *ACM Computing Surveys*, 38(2):#4.

[Zhang and Shasha, 1989] Zhang, K. and Shasha, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal of Computing*, 18(6):1245–1262.