Layout and Language: Challenges for Table Understanding on the Web

Matthew Hurst
WhizBang!Labs
4616 Henry Street
Pittsburgh
PA 15232
USA
matthewhurst@whizbanglabs.com

Abstract

In this paper, we consider the table understanding task and present a catalogue of particular issues that arise when the tables are those found on the web. In addition, we consider what happens when processes commonly associated with web pages are applied to those bearing tables.

1 Table Understanding and the Web

The ubiquity of tables, and their ability to describe relational information in a compact and immediate manner make them attractive targets for automated understanding. Recent research into the automatic location, recognition and understanding of tables has demonstrated the viability of integrating automated table processing systems into larger knowledge management applications ([8]).

However, table understanding is still a relatively novel research area, one whose definition and terminology are still not fixed. It is useful to break the task down into some subtasks, and to consider them in turn with respect to the understanding of tables delivered on the web. Generally, table processing can be conceptualized as consisting of table location; table recognition; functional and structural analysis; and finally interpretation - the extraction of meaningful and unambiguously structured information ([4]). We concentrate on the first two tasks in the following.

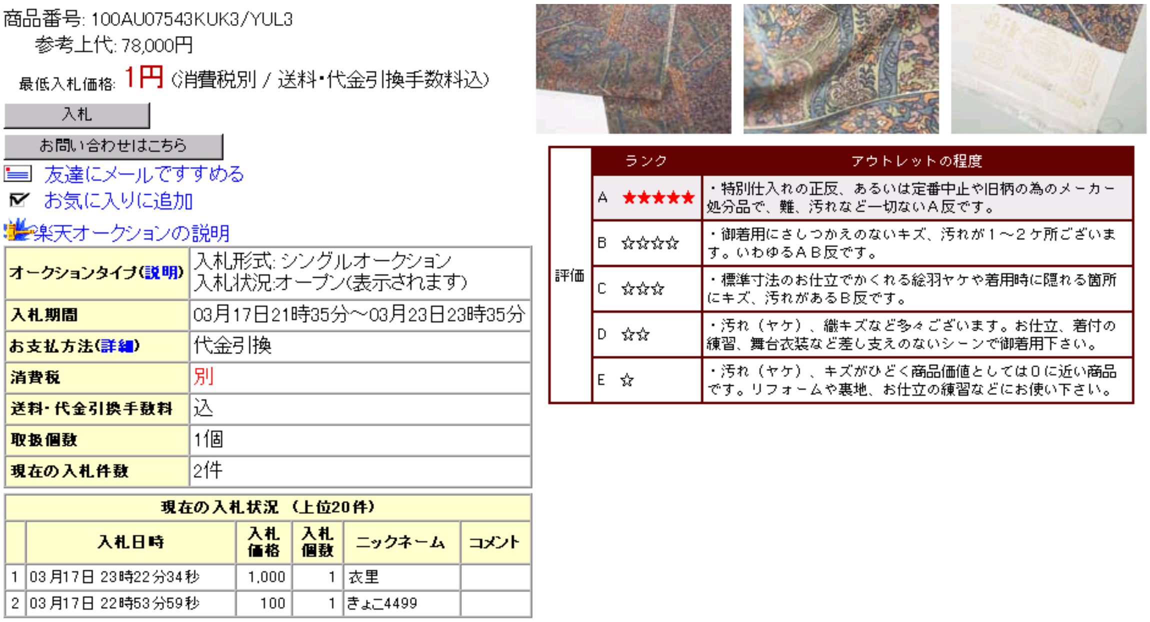
location table location is the processes of spotting tables in documents. Traditionally, this task comes in two basic forms - document image sourced tables ([7], [3]) and electronic text sourced tables including HTML ([1]). The problem is extended to include the spotting of tables in other document encodings such as postscript, pdf, rtf, word, etc. In general, when considering tables on the web, the appropriate HTML tags are exploited (`TABLE`, `TH`, `TD`, etc.). However, this is where we come to the first two distinguishing points.

- the presence of the `TABLE` tag in an HTML document does not necessarily indicate the presence of a table ([1] suggest less than 30 % of HTML `TABLE`s are real tables in one particular domain).
- there are many other ways in which tables may be presented in web delivered documents - plain text (`PRE`), images, mixtures of table specific tags (`TABLE`, etc.) and tags used within the table for their functionality in terms of placing text spatially (`PRE`, `LI`, etc.) - see Figure 1 for an example of such complexities.

The first point requires the creation of accurate classification technology. Given any `TABLE` node in the HTML, the classifier must accept or reject it. Such a classifier may be built either via hand crafted rules ([1]) or using a machine learning approach. Experiments suggest that a machine learning approach using a naive bayse classifier ([9]) based on a feature set describing the set of tags below the potential `TABLE` node in the document tree produces adequate results.

Locating tables encoded in other formats requires technology from other areas. For example, images of tables may be processed by techniques from the document image field ([2]), pre-formatted tables (using the `PRE`) tag may be processed using plain text table methods ([5]). However, the classification problem extends to these cases and individual classifiers must be constructed to make decisions about document elements of each type.

The remaining outstanding issues relate to the mixture of encoding types (e.g. tables built out of `TABLE` nodes and pre-formatted elements), as well as the mixture of encoding purposes (e.g. the use of the HTML `TABLE` to encode surrounding text as well as an embedded table).



商品番号: 100AU07543KUK3/YUL3
参考上代: 78,000円
最低入札価格: 1円 (消費税別 / 送料･代金引換手数料込)

オークションタイプ(説明)	入札形式: シングルオークション 入札状況:オープン(表示されます)
入札期間	03月17日21時35分～03月23日23時35分
お支払方法(詳細)	代金引換
消費税	別
送料･代金引換手数料	込
取扱個数	1個
現在の入札件数	2件

現在の入札状況 （上位20件）

	入札日時	入札価格	入札個数	ニックネーム	コメント
1	03月17日 23時22分34秒	1,000	1	衣里	
2	03月17日 22時53分59秒	100	1	きょこ4499	

	ランク		アウトレットの程度
評価	A	★★★★★	・特別仕入れの正反、あるいは定番中止や旧柄の為のメーカー処分品で、難、汚れなど一切ないＡ反です。
	B	☆☆☆☆	・御着用にさしつかえのないキズ、汚れが１～２ケ所ございます。いわゆるＡＢ反です。
	C	☆☆☆	・標準寸法のお仕立でかくれる絵羽ヤケや着用時に隠れる箇所にキズ、汚れがあるＢ反です。
	D	☆☆	・汚れ（ヤケ）、織キズなど多々ございます。お仕立、着付の練習、舞台衣装など差し支えのないシーンで御着用下さい。
	E	☆	・汚れ（ヤケ）、キズがひどく商品価値としては０に近い商品です。リフォームや裏地、お仕立の練習などにお使い下さい。

Figure 1. A web page using a mixture of HTML tables (on the left) and images of tables (on the right).

recognition table recognition is the task of segmenting the original description of the table into a relative spatial description. In general this task is required when the input is low-level, such as a document image or an electronic text. Clearly, if such tables are found on a web page, the same process is required. Again, given certain assumptions, we can take the marked up tables in a web page to be the logical spatial table. However, there are certain issues that need to be understood in order to account for certain variations:

internal cell structure though tags like TH and TD may be assumed to delimit a single cell in the table, there are cases where other non-table tags are used to provide internal structure in such a way as to associate the cell's contents with those of other cells. A solution would be required to apply a certain amount of recursive processing working into the structure and building a unified abstract table.

split cells in order to gain more control over the distribution of the text in a cell, authors occasionally split the text and place it in two or more adjacent cells. This problem may be accommodated by exploiting linguistic process as described in [6] where the content of the cell can be used to indicate continuity, if any, to other cells.

errors spanning errors occur when the COLSPAN or ROWSPAN values are not correctly calculated. There are two cases. In the first the cell spans beyond the border of the intended table giving the cell incorrect coordinates. In the second, the span of the cell does not communicate the correct meaning of the cell. For example, a cell that is intended to span three cells below it spans only one leading to ambiguity. The first type of problem may be repaired by some form of normalization, whereas the second requires intelligent processing in order to distinguish the following two cases:

	Number of	
Dogs	Cats	Horses

	Date of	
Name	Birth	Address

This is a common problem deriving from the use of HTML as a tool to position document elements on the page rather than a means to encode any part of the logical structure of the document.

omissions the HTML table markup language does not give any reliable control for inserting 'pauses' into tables, e.g. partial line-art, or vertical space. Consequently, empty rows and columns may be inserted. The system must distinguish such intended cells from errors or missing data.

constraints HTML provides a means to encode a table for presentation. Essentially, HTML is a set of operations guiding a rendering algorithm. The TABLE object, and associated elements, are not constrained by the syntax of HTML to encode only those tables that may be correctly rendered by a tree walking rendering algorithm ([11]).

reconstructing HTML one of the first obstacles that any system dealing with documents on the web has to deal with is broken HTML. This often requires the insertion of missing close tags as well as the reordering of incorrectly nested elements and the insertion of missing elements. In the case of tables, using a tool such as Tidy ([10]) may often result at a compromise between the requirements of the HTML specification and the intentions of the author which delivers an unlikely if not incorrect table.

The subsequent tasks (functional analysis, structural analysis, interpretation) are, at a certain level, equivalent for applications dealing with documents from any source, modulo the points made in the remainder of this paper concerning the context in which the table appears.

2 Evaluation

As with any novel field, the table understanding research community is still formulating approaches to the evaluation of their systems. Evaluation requires a precise description of the tasks, as well as descriptions of what is considered ‘the right answer’. Another key aspect of evaluation is the creation of representative corpora. This is one area in which table research is greatly lacking. It is important to have some understanding of the distribution of phenomena in the domain. In other words, we want to know what type of tables occur and how often. Satisfying this requires, formally, the adoption of a model of tables or, informally, the adoption of terminology to describe certain observable features or combinations of features.

Table understanding on the web faces another challenge in terms of evaluation - the potential presence of automatically created tabular data. As the web becomes more interactive, we are seeing many query / response systems returning results in the form of tabulated data. For example, querying an online book store will provide a set of satisfying results listed as a simple table per hit. Such technologies ensure that there is no realistic way in which we can provide a distribution of tables in any general sense.

One advantage that the web offers in this area is the access to certain cites which contain a large set of pages containing similar tables describing related information. For example, product description sites for large corporations often include specification and feature tables as well as product comparison tables. Such sites can be mined for large sample sets where processing can be carried out on a reasonably restricted domain.

3 Context

Research has suggested that the context in which the table occurs provides many useful resources for developing an understanding of the table ([4]). Due to the hyper-linked nature of the web, however, there is potential for tables to be isolated from the document it logically occurs in. Giving a web page including a table to a system might remove the important information that may be found in the remainder of the document.

Conversely, hyper-linking permits the arbitrary linking to pages created by different authors, using different terminology and even different table conventions. It has been suggested that tables written in different languages, or by authors with different native languages include certain specific variations . If the table processing system implements these local assumptions inflexibly, then the interpretation of tables based on a different set of assumptions may not be achieved. For example, the following structure is common in Japanese tables:

General Term		
	Sub Term A	Sub Term B
data$_G$	data$_A$	data$_B$

though not in western tables. The same variation can be found between genres and domains of discourse, suggesting a level of specialization for the implemented system.

4 Common Web Applications and Tables

For tables to be included in the web as a whole, we must consider the set of operations expected of web based documents and how the table may be accommodated.

search searches on the web are carried out by inputting of simple search terms, and the retrieval and ranking of hits. In general, the distance between and order of search terms is assumed to be significant. Underlying this assumption is the simple fact that words that modify each other phrasally (in English) occur in close proximity. However, searching for particular combinations of concepts in tables cannot rely on such assumptions. Raw HTML encodes text in a very transparent manner, however tables are encoded in a way that neither reveals the semantic relationship between constituents, nor which directly describes the spatial relationship between them. Consequently, search directly targeted at tables, or searches that may also include tables, require the ability to recognize such distinctions.

clustering, classification in order to provide added value, many search sites offer some form of retrieval of similar documents. Underlying this is technology to identify clusters of documents. This clustering is often performed only via the inspection of words with no recourse to document structure. In order to accommodate tables, we must consider what makes two tables similar, and, what makes a table similar to a document.

summarization delivering search results often requires the delivery of some form of brief description of the page - either a summary, or the quoting of a significant passage. If the document is a table, how can such a fragment be found, and how should it be delivered. Simply outputting the contents of the cells as they are discovered in the HTML is almost meaningless.

translation translation is becoming another popular, if not yet mature, feature of web portal services. It might be assumed that the translation of tables is simply the translation of the linguistic fragments found in the cells. There are two points to consider: firstly, linguistic context is important to good translation - the fragment of text in a cell is often better understood in association with the text discussing it; secondly, the language used in table cells, similarly to that used in section headings, headlines, titles and so on, is often truncated or simplified in some manner.

In general, all the issues contained in this section are table-complete in that they require a full parsing of the structure of the table as well as, potentially, its interpretation in order to deliver complete results.

5 Conclusions

This paper has attempted to highlight some of the challenges that are faced if the complexities of tables are to be fully utilized and understood to ensure that their contents are made accessible in the same transparent way as the contents contained in plain text on the web. Essentially, the impoverished HTML table encoding, abuse of the HTML table tag set, the presence of images, plain text and other issues suggest that the full exploitation of information contained in tables on the web requires a reasonably complex table understanding module in all applications associated with web pages, both server side and client side.

The role of document analysis in web content extraction in the case of tables is in general similar to that for normal documents. However, there is a requirement that systems be flexible enough to deal with all forms of table encoding and their idiosyncrasies, errors and conventions. The assumption in web content understanding is that HTML is a logical encoding of the document, and as such is sufficient for the task. It is not clear if this assumption is correct in general. Specifically for tables, there are problems due to the inability of HTML (or any tree-like document encoding) to capture the logical structure of the table.

The state of the art of table understanding suggest that it is possible to interpret tables from a specific domain with reasonable accuracy due to the presence of domain knowledge. However, open table understanding is not yet possible, and it is this capability that is required, at least to the stage of deriving the logical structure of the table, if the common web applications are going to be made availably transparently for tables.

References

[1] H.-H. Chen, S.-C. Tsai, and J.-H. Tsai. Mining tables from large scale html texts. In *The 18th International Conference on Computational Linguistics*, Saarbrucken, Germany, July 2000.

[2] D. Dori, D. Doerman, C. Shin, R. Haralick, I. Phillips, M. Buchman, and D. Ross. *Handbook on Optical Character Recognition and Document Image Analysis*, chapter The Representation of Document Structure: a Generic Object-Process Analysis. World Scientific Publishing Company, 1996.

[3] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong. Medium-independent table detection. In *Document Recognition and Retrieval VII*, pages 23 – 28, San Jose, California, USA, 2000. SPIE.

[4] M. Hurst. *The Interpretation of Tables in Texts*. PhD thesis, University of Edinburgh, School of Cognitive Science, Informatics, University of Edinburgh, 2000.

[5] M. Hurst. Layout and language: An efficient algorithm for text block detection based on spatial and linguistic evidence. In *Document Recognition and Retrieval VIII*. SPIE, 2001.

[6] M. Hurst and T. Nasukawa. Layout and language: Integrating spatial and linguistic knowledge for layout understanding tasks. In *Proceedings of the 18th International Conference on Computational Linguistics*. ICCL, July 2000.

[7] T. G. Kieninger and B. Strieder. T-recs table recognition and validation approach. In *AAAI Fall Symposium on Using Layout for the Generation, Understanding and Retrieval of Documents*. AAAI, 1999.

[8] D. Lopresti and G. Nagy. Automated table processing: An (opinionated) survey. In *The Third IAPR International Workshop on Graphics Recognition(GREC '99)*, 1999.

[9] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[10] D. Raggett. Html tidy. http://www.w3.org/People/Raggett/tidy.

[11] D. Raggett. Html 4.01 speification. Technical Report http://www.w3.org/TR/html401, W3C, 1999.