Evaluation of Information Extraction Techniques to Label Extracted Data from e-Commerce Web Pages

Neil Anderson
School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast
Belfast BT7 1NN, UK
n.anderson@qub.ac.uk

Jun Hong
School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast
Belfast BT7 1NN, UK
j.hong@qub.ac.uk

ABSTRACT

Automatically determining and assigning shared and meaningful text labels to data extracted from an e-Commerce web page is a challenging problem. An e-Commerce web page can display a list of data records, each of which can contain a combination of data items (e.g. product name and price) and explicit labels, which describe some of these data items.

Recent advances in extraction techniques have made it much easier to precisely extract individual data items and labels from a web page, however, there are two open problems: 1. assigning an explicit label to a data item, and 2. determining labels for the remaining data items. Furthermore, improvements in the availability and coverage of vocabularies, especially in the context of e-Commerce web sites, means that we now have access to a bank of relevant, meaningful and shared labels which can be assigned to extracted data items.

However, there is a need for a technique which will take as input a set of extracted data items and assign automatically to them the most relevant and meaningful labels from a shared vocabulary. We observe that the Information Extraction (IE) community has developed a great number of techniques which solve problems similar to our own. In this work-in-progress paper we propose our intention to theoretically and experimentally evaluate different IE techniques to ascertain which is most suitable to solve this problem.

General Terms

Algorithms, Design, Experimentation

Keywords

Data item labelling, vision-based extraction

1. INTRODUCTION & MOTIVATIONS

Web sites that rely on structured databases for their content are ubiquitous. Users retrieve information from these databases by submitting HTML query forms. Query results are displayed on a web page, but in a proprietary presentation format, dictated by the web site designer. We call these pages Query Result Pages. Automatic data extraction is the process of extracting automatically a set of data records and the data items that the records contain, from a Query Result Page. Such structured data can then be integrated with data from other data sources and presented to the user in a single cohesive view in response to their query. For instance, there is great commercial demand for comparison shopping search engines. A user may wish to buy a book: a comparison shopping search engine can extract data from many different online stores, integrate the data and display it to the user. Other practical applications include flight and hotel booking sites, financial product comparisons, property sales and rentals. A Query Result Page is designed for a human to read rather than a computer to process, thus there is no standard way to extract automatically structured data from the page.

Figure 1 illustrates a typical Query Result Page from waterstones.com. On this page each book is presented as a data record, which contains a set of elements. These elements are either data items or labels corresponding to data items. For example, the book prices, '£7.99' and '£5.59' are examples of data items in each record while the text 'Format' and 'Published' are examples of labels in each data record.

In our previous work we have solved a number of data extraction problems.

In [2] we present a visual approach to data record extraction, which identifies the boundaries of each record on a Query Result Page. As illustrated by Figure 2, the boundary of each data record is outlined with a red bounding box, while the elements contained in each data record are outlined with blue bounding boxes.

In [1] we present a learning classifier-based approach to semantic grouping, which aligns the elements into similarity groups. Our approach also classifies each similarity group as either a group of data items or a group of labels. As illustrated by Figure 3, colour coding can be used to indicate membership of a similarity group, for example, the 'price' data item in each data record is outlined in pink, while the 'product names' are outlined in green. Data items are outlined with a solid line, while labels are outlined with a dashed line.

Following on from our previous work, a number of problems remain open. First, most data items on a typical Query Result Page are labelled implicitly. This means that the web

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579703.

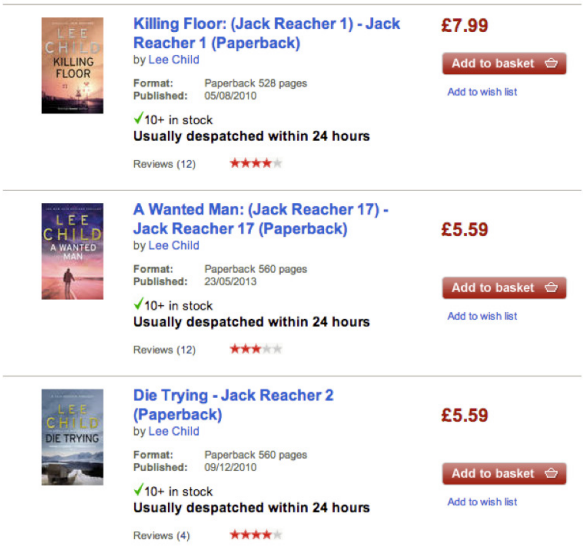

Figure 1: A typical Query Result Page from waterstones.com

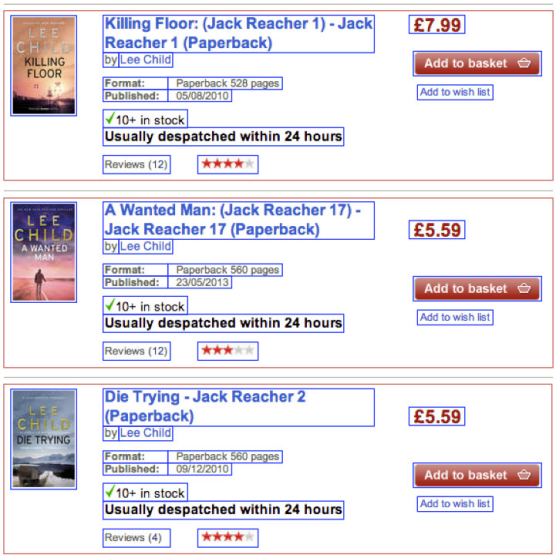

Figure 2: A Query Result Page from waterstones.com with visual blocks highlighted.

site designer has not provided an explicit text label for the data item. Instead, the designer has assumed that the reader of the page can easily infer the semantic meaning of the unlabelled data items based on their visual context and their own prior knowledge. Our approach must solve this problem by inferring a semantic meaning for these data items.

Second, occasionally some of the data items on a Query Result Page are labelled explicitly. A naïve approach would attempt to use the provided labels verbatim. However, there are multiple issues with this strategy:

1) There is no standard convention for data item and label association. The label could be above or below, left or right, aligned or not aligned, proximal or not proximal to the corresponding data item. One label could correspond to more than one data item, or one data item could have more than one corresponding label.

2) There is no standard set of labels in shared use between different web sites (even considering web sites in the same domain of interest). For example, one web site may use the label 'List Price' to indicate the full price of a product, while a similar web site selling the same product at a discount might use the label 'Was' to indicate the full price.

These problems make it very difficult to achieve accurate integration of data items from different web sites. Accordingly, we are motived to develop techniques which will assign a meaningful label to each similarity group of data items on a Query Result Page. Furthermore, the label should be shared and consistent. This means that two similarity groups from different web sites that share the same semantic contents should be assigned the same label.

Fortunately, 'The Big Three' search engines (Google, Bing and Yahoo!), have recently developed a common set of set of schemas for structured data markup on webpages [7]. These schemas, called *Schema.org* (further description of *Schema.org* is provided in Section 4), are intended to describe entities on the web, such as products, people, organisations, creative works and much more. The schemas are organised into types and properties. Types are things like Book, Person and Product. Each type has many properties, for example the Product type includes the properties: depth, width, weight, description and brand. For our work, the properties of a type are, effectively, a set of labels. Moreover, since *Schema.org* is supported by The Big Three, then these labels can said to be shared.

Our intention is use the properties of the types in *Schema.org* to help us label the data items that we have been able to extract from a Query Result Page.

To do so, we could write manually a set of rules to assign each extracted data item to a property in the schemas. Such an approach might work well for a few chosen data items, but would quickly prove brittle and un-workable on a larger scale. Another option would be to use a simple classifier-based approach to reconcile data items with schema properties. This would be more flexible than manual rules, however, our experience of classifier-based approaches is limited. Therefore, to solve this problem effectively and efficiently, we cast it as an IE problem.

The rest of this paper is organised as follows. We explore, in Section 2, a number of different IE approaches that have been used to solve problems similar to our own. In Section 3, we first describe a visual block model which captures all the visual, spatial and content features of each element on a Query Result Page. Second, we explain how we use these features to build a feature vector which describes the characteristics of an element. In Section 4, we discuss our choice of *Schema.org* as a vocabulary for labelling data items on e-Commerce Query Result Pages. Finally, Section 5 outlines our ongoing work and concludes the paper.

2. INFORMATION EXTRACTION

The IE field is awash with mature, tried and tested Machine Learning-based techniques which are frequently used to solve problems similar to our own. Our goal is to theoretically and experimentally evaluate the effectiveness of a range of different IE techniques. Specifically, we are inter-

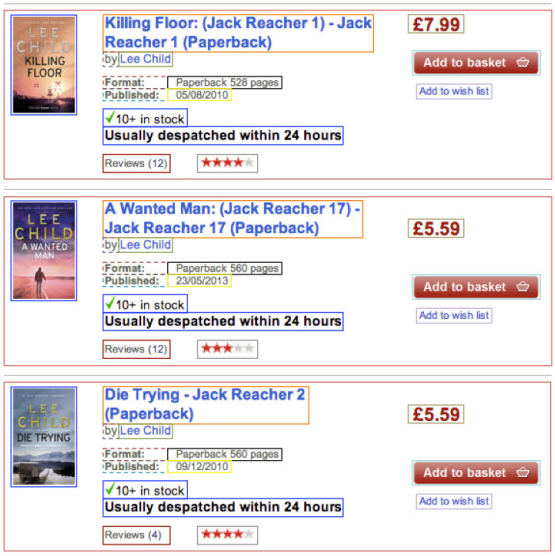

Figure 3: Colour-coding of visual blocks indicates membership of a semantic group.

esting in techniques which are either: 1. classifier-based or 2. Markov model-based.

In the first category, there are a number of classification choices frequently used in IE, including Support Vector Machines (SVM), decision trees and naïve Bayes. Furthermore, in [5, 6], multiple classifiers have been combined effectively to solve an IE problem, in this instance Named Entity Recognition. As detailed in Section 3, our approach can create a feature vector for each data item which can be used by a classification-based (or combination of classifiers) technique.

In the second category, a number of Markov model-based techniques have been shown to be highly effective for sequence labelling tasks. At first it may not seem entirely clear how these techniques could be carried across to solve our problem of labelling seemingly isolated and unlinked data items. However, while the data items may appear to be completely segmented, they are infact intrinsically linked by measures such as proximity and alignment of data items.

Our intention is to represent the relationships between the individual data items in a data record using proximity and alignments measures. We will then attempt to tag each data item using a Markov-based technique. For example, in [11] Hidden Markov Models (HMMs) are used to tag parts of speech. In [8], Conditional Random Fields (CRFs), an alternative to HMMs, are used to build probabilistic models to segment and label sequence data.

We examined two IE techniques, Named Entity Recognition and Part-of-Speech Tagging, to understand if their approaches could be carried over to help us solve our problem.

2.1 Named Entity Recognition

Our goal appears to be aligned to that of Named Entity Recognition (NER). In [12], Named Entities are defined as, "phrases that contain the names of persons, organizations, locations, times, and quantities". Our interest centres on Query Result Pages from e-Commerce domains. These pages typically contain a much richer and more diverse set of types of data items than the Named Entity definition allows. For example, a typical data record could contain the following data items with attributes including: price, delivery price, description, stock, level, photograph, weight and product codes to list a small subset. Furthermore, established resources, such as, *YAGO*[10] and *Freebase* [4], make use of a limited selection of hierarchies of named entity types. These are not of great relevance to our task as they are not specific enough to e-Commerce products. These distinctions effectively bar us from using an NER approach straight off the shelf, however, other relevant IE techniques such Part-of-Speech Tagging look promising.

2.2 Part-of-Speech Tagging

Part-of-Speech Tagging (POST) is the process of attaching tags, each corresponding to a particular grammatical part of speech, to the text contained in a given corpus. The choice of tag is based on both the definition of the word in the corpus, as well as its context, i.e. a given word's relationship with adjacent words in the phrase or sentence. For instance, in [3], the relationship between adjacent words on both sides of the word to be tagged form the context.

While it is not entry clear if any of the existing approaches for part-of-Speech Tagging could be directly carried over to help us achieve our goal of data item labelling, we have learnt from POST that when labelling a data item, we must consider both the description of the data item (akin to the definition of a word) as well as the context of the data item in relation to other data items in the same data record.

At this stage, we do not know which of the above techniques are best placed to solve our specific problem: our goal is to experimentally select the most suitable IE technique.

3. VISUAL APPROACH

Our approach to alignment and classification of the elements of data records takes as input a *Visual Block Model (VBM)* of a Query Result Page. The VBM of a Query Result Page is a product of the tag tree and the Cascading Style Sheet (CSS) of the page. A layout engine generates a visual block for each node in the tag tree, according to the instructions contained in the CSS. This process, called rendering, draws a rectangular box around the minimum boundary of each visible node on the page. We refer to each rectangular box as a visual block. The position of each visual block is represented by its four borders in the four directions on the two-dimensional plane. The plane has its origin at the top-left of the page, with the x-axis running from left to right and y-axis running from top to bottom.

As we are only interested in the data records and their contents in this work, we employ the *rExtractor* algorithm in [2] to identify the boundaries of each data record on the Query Result Page. Once each element in a data record has been represented as a visual block, our approach employs the algorithm in [1] to classify each visual block as either a data item or a label. Finally, for each data item in a data record, our approach will create a data item definition and a data item context.

3.1 Create a Data Item Definition

We employ the algorithms as described, in detail, in [2] to create a feature vector. Automatic feature extraction employs a collection of algorithms to extract a feature vector

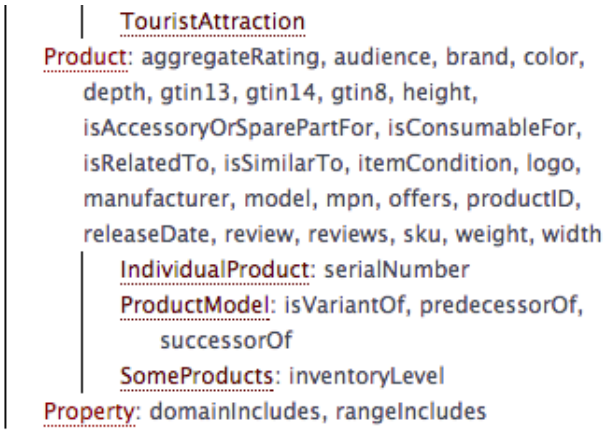

Figure 4: A snippet of *Schema.org*, showing the product type and properties.

to define, or describe, each data item. The goal of each algorithm is to extract a particular type of feature from a visual block representing a data item. Each feature describes a visual, structural or content characteristic of the block. The characteristic include: visual, identity, formatting and punctuation features.

3.2 Create a Data Item Context

A data record typically contains a number of data items. We call the relationship between one data item and the other data items in the same data record the *context* of a data item. For each data item in a record, it is proposed that our approach will create a feature vector to describe its relationship (measured in terms of alignment and proximity) with each of the other data items in the same data record.

4. SCHEMA.ORG

Schema.org, a snippet of which is shown in Figure 4, is a shared markup vocabulary created and recognised by major search providers. When it was launched in 2011 it contained a number of types and properties that made it effective for marking up content from e-Commerce web sites. In late 2012, *GoodRelations* [9], which is a web vocabulary for e-Commerce, was fully integrated into the *Schema.org* vocabulary. This update enhances our confidence in Schema.org, first because it increases the granularity with which we can potentially label data items on e-Commerce web pages and second because it demonstrates that *Schema.org* is a relevant vocabulary, which is frequently updated to ensure its is reflective and representative of current web content and trends.

5. CONCLUSIONS & FUTURE WORK

This paper presents an investigation into the merits of applying IE techniques to the problem of labelling extracted data. We found that there are number of similarities between our problem and those problems solved by tried and tested IE techniques. In short there is merit in the continued investigation of the application of IE techniques to solve our labelling problem.

We plan to continue this work to: 1. theoretically and experimentally evaluate, in the context of our labelling problem, the IE techniques we describe in this paper and 2. implement the most effective IE technique to solve our problem of automatically labelling data items extracted from an e-Commerce web page.

6. REFERENCES

[1] N. Anderson and J. Hong. *A Learning Classifier-based Approach to Aligning Data Items and Labels*, pages 282–291. Springer, 2013.

[2] N. Anderson and J. Hong. *Visually Extracting Data Records from Query Result Pages*, pages 392–403. Springer, 2013.

[3] M. Banko and R. C. Moore. Part of speech tagging in context. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[4] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA, 2008. ACM.

[5] X. Carreras, L. Màrquez, and L. Padró. Named entity extraction using adaboost. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[6] R. Florian, A. Ittycheriah, H. Jing, and T. Zhang. Named entity recognition through classifier combination. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 168–171, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[7] http://tinyurl.com/pe5e87b. Introducing schema.org: Search engines come together for a richer web, 2011.

[8] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[9] G. Relations. Good relations: Web vocabulary for e-commerce, 2014.

[10] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.

[11] S. M. Thede and M. P. Harper. A second-order hidden markov model for part-of-speech tagging. In *In Proceedings of the 37th Annual Meeting of the ACL*, pages 175–182, 1999.

[12] E. F. Tjong Kim Sang. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.