

PDFs, cognitive computing and layout analysis

Introduction

In this paper we will be looking at the role of layout analysis. Where it is useful, where not – particularly with cognitive computing in mind.

Why are PDFs difficult to handle by cognitive computing?

Documents can have complex structures made up of successive chapters or sections over many pages and having regions with various different types of information: paragraphs, words, letters, graphics, images, tables, logos, equations, columns...

A human eye can immediately make sense of a page and automatically organise the information into letters, words, paragraphs, tables... A person doesn't need to be told where a word starts and ends. The brain re-creates the layout without the need of additional information such as word delimiters & boundaries.

Computers on the other hand are not good at this task and need to be told where the words, paragraphs, tables and other boundaries are. This information is usually defined in tags – for instance the tag <p> in html is used to delimit paragraphs.

Unfortunately many documents - including PDFs - have virtually no structural information.

PDF files have no underlying tags to delimit words, paragraphs or tables. A PDF at a low level can be seen as an assembly of letters and pixels with no concept of words, paragraphs, tables, diagrams, maps, etc. The PDF format was designed for printing - with no need for structural information or content labeling.

This lack of structure explains why it isn't possible to copy and paste a table into excel, and why one cannot directly edit and change a PDF file.

What is meant by 'cognitive computing'?

It's a broad term, and generally understood to be part of artificial intelligence (AI) that deals with 'understanding' written or spoken language. Some examples of cognitive computing might help

sentiment analysis, e.g.

- trying to assess attitudes towards a consumer product from thousands of Tweets

article summarisation – creating a paragraph that gives an overview of a long document

drug discovery - 'mining' lifescience reports for new connections

speech recognition – e.g. giving instructions via Apple's Siri or Microsoft's Cortana

In all these, a computer is doing a complex language task.

What's made cognitive computing happen now?

The ongoing revolution in the field of Artificial Intelligence was made possible by the combination of two factors :

- 1) the access to large corpora of text and data, and
- 2) the ability of computers to learn without being explicitly programmed (Machine Learning).

A scanned document (as an image or as a PDF) has even less layout information than a native PDF – it doesn't even differentiate letter pixels from image pixels. An OCR system is needed to recognise a group of pixels as representing a specific letter.

The lack of structure can be fixed by processing PDFs through complex algorithms that attempt to identify the various regions, the flow of text, etc.

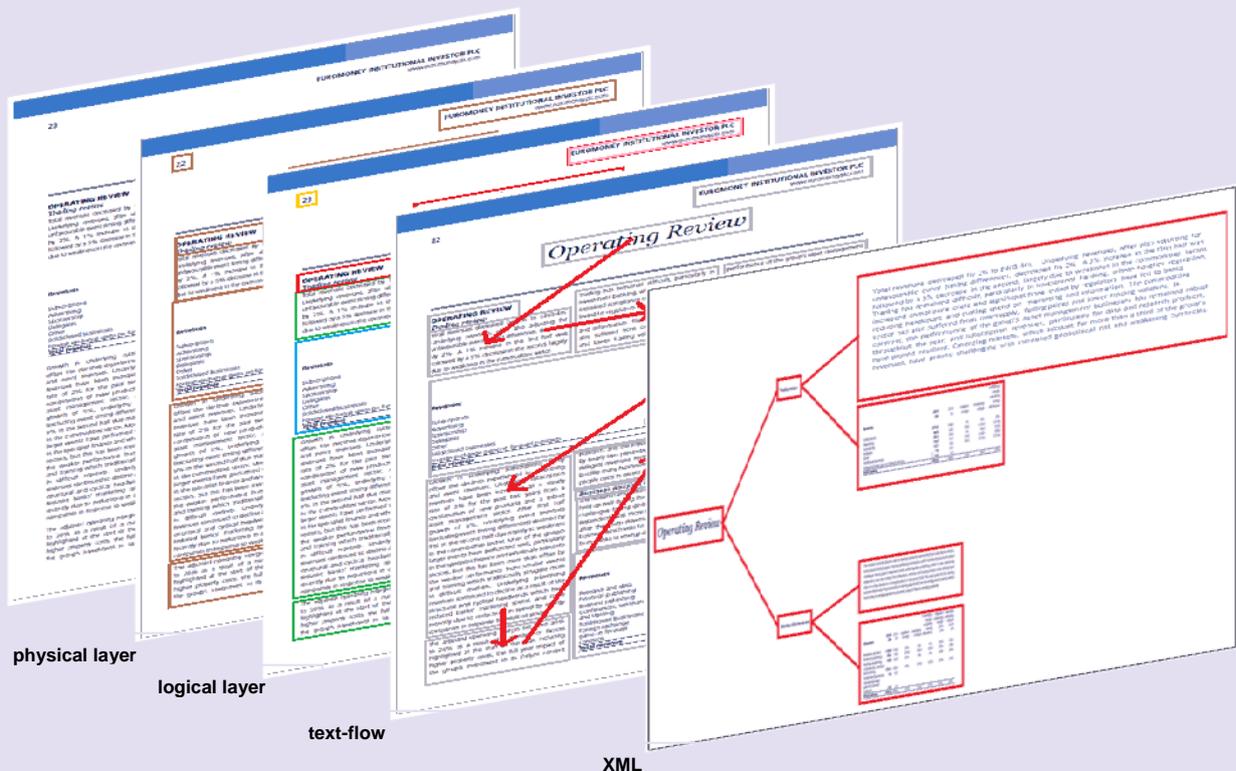
Layout analysis

The identification of all the regions on a page is done through a process known as 'document layout analysis'.

Document layout analysis first divides up the document into regions with homogeneous content and secondly assigns a meaning to the regions. The 'dividing up' step is called 'physical layout analysis', and is used to identify the geometric page structure. It's essentially a visual process – not reliant on the meaning of the content.

Taking the regions and labeling them (as titles, captions, footnotes, sections etc.) is the scope of 'logical layout analysis' and is essentially semantic – it relies on analysis of language. Document layout analysis is the combination of geometric and logical labelling.

On top of these two processes one can add a text-flow analysis layer and a tree-like representation layer of the document content. All this information can easily be kept in a single XML file.



In the figure above, the physical layout analysis is shown as the second layer and the semantic labelling as the third layer. The fourth layer shows the text-flow. And the fifth shows the representation as XML.

So when is this useful?

Examples and applications

We'd like to give several examples to give an indication of when layout analysis makes content extraction and cognitive computing easier or more viable. We'll start by briefly mentioning some cases where it is NOT useful.

Layout analysis NOT necessary

The key here is either to have text that's simple, or where it doesn't matter greatly if you get it wrong.

Example 1: [extracting a standard piece of data from a document](#) – for example, extracting the names of Chief Executives from Annual Reports

Our established businesses are progressing in line with our expectations and we expect the Group to deliver good growth in 2016. In the UK, we expect our business to be stable with growth being driven by our international businesses. Increased customer acquisition investment and a weaker Euro are expected to reduce profits in France, however this will be more than offset by strong growth in the USA and Spain. We plan to invest around £6m in New relation to innovation initiatives.

Richard Harpin
Chief Executive
19 May 2015

We are at our best when we are **completely focused on our customers**. My plan is to keep things simple by putting them at the heart of M&S – every decision starts with them.

STEVE ROWE CHIEF EXECUTIVE

You can find the names automatically by using a well-designed regular expression that will pick out the capitalised words that come before 'Chief Executive', 'CEO', etc. You don't need to know the layout of the pages.

Example 2: [enterprise search](#)

If you are building a search engine to find internal documents, emails, etc, you need to index the text of the documents. It doesn't matter greatly if the text order is slightly wrong or if some spurious details are included (page numbers, footnotes, graph labels, etc). So you don't need to know the layout of the pages.

Example 3: [sentiment analysis of simple text](#)

If, for example, your source text has simple layout – like millions of Tweets or emails – then your analysis can focus on the text. You don't need to know the layout of the pages.

Layout analysis NOT necessary - *continued*

Example 4: ingesting invoices

Many large accounting systems take in invoices as PDFs (native or scanned). The content is extracted into the accounts package because the invoices from each supplier will look similar every time. The data can be extracted using a template which is built for each invoice type.

INVOICE

PLEASE REMIT TO: Page 1 of 1
 DJO, LLC
 PO BOX 650777
 DALLAS TX 75285

Standard positions on page

Invoice No.	Invoice Date
17847897	26/01/2017
Purchase Order	Customer No.
PO/17/001	327069
Order No.	Order Date
6632148	22/01/2017

sales1@natsaivitalhealth.com

Bill To: 327069 NATSAI VITALHEALTH PTE LTD
 NO. 8, ANGKLONG LANE
 #01-03, FABER GARDEN
 SINGAPORE 579981
 Singapore

INT Ship To: 327069 NATSAI VITALHEALTH PTE LTD
 1418, 14TH FLOOR, KOMPLEKS
 SELANGOR
 50000 JALAN SULTAN, KUALA LUMPUR,
 MALAYSIA
 NATSAI VITALHEALTH (M) SDN BHD
 50000
 Malaysia

Thank you for your order. We appreciate your business.

Ship Date	26/01/2017	Ship Via	TBD EXPORT	Terms	Net 30 Days	Due Date	25/02/2017
-----------	------------	----------	------------	-------	-------------	----------	------------

Item No.	Qty. Ordered	Qty. BO	Qty. Ship	Unit	Unit Amt.	Tax	Total Amount
1 83162 TUBE COUPLER TRT800	1	0	1	EA	45.50	N	45.50

Subtotal	45.50
Tax Total	0.00
Invoice Total	45.50

Although at first sight it appears that this would be a good application for layout analysis and cognitive computing, in practice the problems are addressed differently – because of the repeat nature of the documents. It doesn't need either layout analysis or AI.

Layout and text-flow analysis necessary

Example 5: extracting specific text from a document

5). First, here is an example where, to extract the text from each section of the document, the computer needs to understand both 'sections' and 'columns' as well as the text-flow. In other words you need layout analysis.

CF Woodford Income Focus Fund, C Sterling Income, a fund within CF Woodford Investment Funds II (ISIN: GB00BD9X6V34)
 The fund is managed by Capita Financial Managers Limited, part of the Asset Services Division of Capita plc.

Objectives and investment policy section

Objective
 The fund aims to provide a high level of income together with capital growth.

Investment Policy
 The fund will invest predominantly in shares of companies listed in the UK and overseas with a focus on investments that provide dividends. The fund will be invested in a concentrated portfolio of securities.

Essential features of the fund:

- The fund has the discretion to invest in a range of investments as described above with no need to adhere to a particular benchmark.
- The fund will seek to provide an income of 5p per share per annum. There is no guarantee any specific level of dividend or yield will be achieved over any given time period.

text flow

- You can buy and sell shares in the fund every business day.
- The fund aims to distribute available income every quarter.
- Derivatives are used for investment purposes and to manage the risk profile of the fund.

Recommendation: This fund may not be appropriate for investors who plan to withdraw their money in the shorter term (e.g. less than 3-5 years).

Risk and reward profile

← Typically lower rewards / Lower risk | Typically higher rewards / Higher risk →

1 2 3 4 **5** 6 7

- Counterparty Risk: As the fund may enter into derivative agreements there is a risk that other parties may fail to meet their obligations. This may lead to delays in receiving amounts due to the fund, receiving less than is due or receiving nothing.
- Financial Techniques Impact: The fund invests in derivatives. A relatively small movement in the value of the derivative's

Layout and flow analysis necessary, *continued*

Example 6: *textual analysis of a document*

A cognitive analysis of a document will be much more effective if it can be done on a clean version of the document where all the text in charts & tables, footnotes, page headers, captions... has been removed.

This is necessary to avoid:

- introducing unwanted text such as axis labels, captions, tables headers, etc. in the natural flow of a document.
- interfering with cognitive algorithms; for instance, a search for a paragraph section with the phrase 'profit & loss' risks being pulled towards table regions that are likely to include some headers with the same words
- interfering with the statistical distribution of words; most algorithms rely on the word counts for indexing, categorization, searching purposes. The 'pollution' of the word distribution by unwanted regions will affect the efficiency of these algorithms

The example below shows how the charts, the table and their captions and notes (artificially given shaded backgrounds) need to be removed to make sense of the core text.

Exhibit 5
77% Renewal Rates Are Positive, but Non-Renewals Are Going to Free Solutions or Turning off Security Software versus Moving to another Paid Solution

Source: Morgan Stanley Research, Alphawise

Exhibit 6
Vendor Relationship Sustains Through New PC Purchases Less than 50% of the Time

Source: Morgan Stanley Research, Alphawise

Increasing Importance of Distribution

Distribution is now becoming increasingly important as differentiation in consumer security offerings becomes less apparent. Our survey results show distribution has come to trump performance or feature/functionality in consumer purchasing habits for endpoint security.

Free trials was the most cited number 1 reason for choosing a consumer security product. At the same time, independent lab tests fail to show significant differentiation between free and paid vendors in terms of product performance or effectiveness — and none of the products are 100% effective. Bottom line,

it's difficult for the consumer to readily identify a clear winner that consistently beats peers on performance and functionality and therefore it's easy for the consumer to go with the default choice on their PC.

Our survey shows that consumers most often choose to go with the vendor offered with their PC purchase with 26% of respondents citing a free trial (either from an OEM, ISP, or other, such as a bank) as the number one reason for choosing their current consumer security software. Richness of features and functionality ranked low, with only 3% citing it as the number one reason for choosing their software.

Exhibit 7
Free Trials are the Most Often Cited #1 Reason for Choosing Consumer Security Software

Source: Morgan Stanley Research, Alphawise. Note: Free trial responses include free trials from OEMs, ISPs, and others.

Only 47% of consumers plan to continue using their current security vendor after their next PC purchase, while 27% will use what ships on the PC, so distribution deals likely remain very competitive going forward. At the same time our survey suggests ~5-7% of the market is moving to free solutions annually, which could pressure renewal rates and conversion rates going forward.

How PC OEM distribution works

PC OEMs are the primary distribution channel for major consumer software security vendors like SYMC and MFE, contributing 60-70% of new customers, by our est. These partnerships are likely a source of high incremental profit margin for PC OEM manufacturers and their strategic importance to software companies gives advantages to PC OEMs in deal negotiations, which has caused pricing to erode — impacting consumer margins longer term. While MFE was more aggressive in 2008 forming partnerships, SYMC has re-asserted itself recently, announcing new partnerships with Samsung and Fujitsu on their last earnings call.

Layout necessary

Example 7: extracting data from tables

When wanting to make sense of the content in tables, layout is vital.

The identification of the table boundaries, its columns, rows, headers... is a very complex layout analysis sub-task.

The cognitive computing can then be used to answer questions like “was industrial production up or down in 2012?”

-on-year decline in turnover on the euro area market and the domestic market slowed over the first four months of the year.

was up just under 2%, ceasing the negative developments in the two previous months. The year-on-year decline in turnover diminished to just under 1%. After recording a sharp decline in the final quarter of last year, this year administrative and support service activities are recovering particularly well for the moment.

Economic Activity	2010	2011	2012	12 m. to Apr-13	2013 Mar.	2013 Apr.	2013 Apr.
				<i>y-o-y in %</i>			++
Industrial production: - total *	7.0	2.0	-0.7	-0.9	-3.3	-1.1	2.0
- manufacturing	7.4	1.8	-1.9	-2.2	-5.6	-1.4	2.0
Construction: - total **	-16.9	-24.8	-16.8	-18.9	-31.7	-19.3	-6.4
- buildings	-14.0	-39.7	-17.3	-26.0	-50.1	-37.2	-2.8
- civil engineering	-19.0	-15.3	-16.6	-13.8	-13.4	-6.7	-6.6
Trade (volume turnover)							
Total retail trade	-0.3	1.7	-2.3	-3.7	-5.9	-3.7	-1.5
Retail trade except automotive fuel	-1.6	-2.2	-4.7	-5.1	-6.3	-1.4	-0.6
- food, beverages, tobacco	-1.6	-2.9	-4.8	-4.3	-3.3	-3.1	-0.8
- non-food (except automotive fuel)	-1.6	-1.7	-4.9	-5.6	-9.2	-0.2	-0.4
Retail trade and repair of motor vehicles	11.9	7.6	-5.3	-5.6	-4.9	4.7	1.0
Private sector services *** +	6.3	3.2	-2.3	-2.6	-4.0	-0.8	0.2
Transport and storage +	19.6	8.0	0.7	-0.4	-5.2	0.3	0.0

Sources: SORS, Eurostat, Bank of Slovenia calculations.

Notes: Data are working days adjusted.

* Volume of industrial production. ** Real value of construction put in place. *** Excluding trade and financial services. + Nominal turnover.

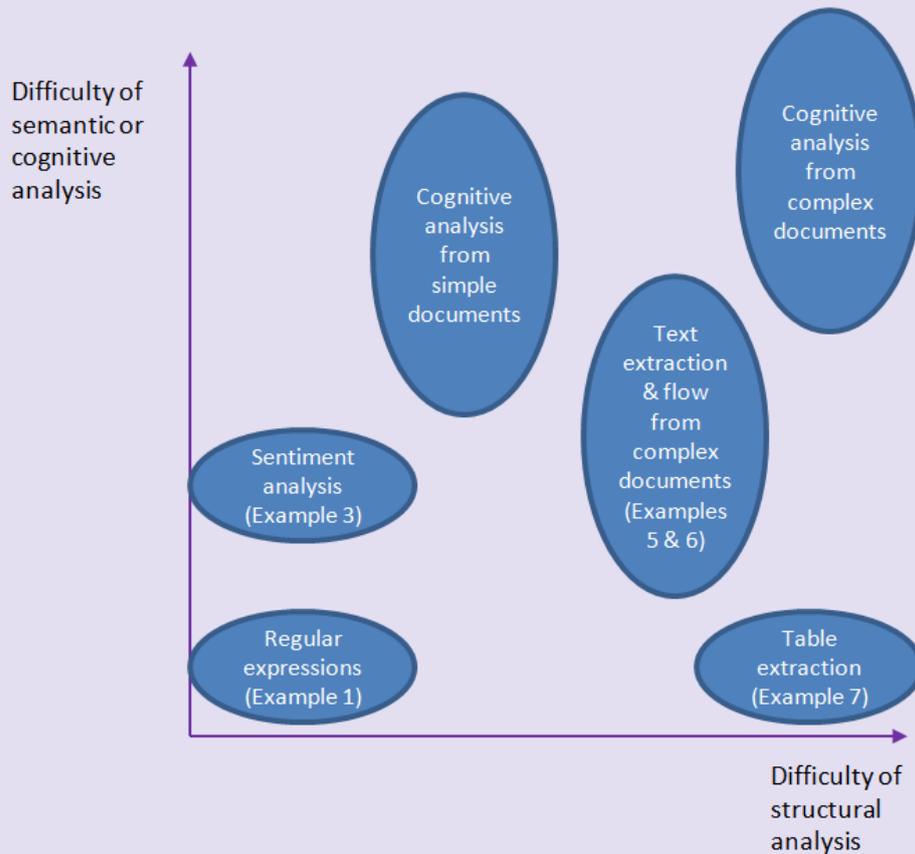
++: 3-month moving average compared to the corresponding average 3 months earlier. Data are seasonally and working days adjusted (except for construction where data are seasonally adjusted).

Conclusion

We contend that layout analysis is most useful when:

- the text for cognitive analysis is interspersed by graphics and tables
- key content is present in tables – which vary from one document to another
- text-flow is important

Finally, when looking at any new project, it's worth considering the following technical matrix:



In other words, it's useful to think of the feasibility of the cognitive component and the feasibility of the structural component independently.