SpendInsight: some remarks on deploying an intelligent spend-analysis system

Richard W. Barraclough
@UK PLC, 5 Jupiter House, Calleva Park
Aldermaston, Reading, Berkshire, RG7 8NN.
richard.barraclough@ukplc.net

J. Mark Bishop & Sebastian Danicic
Dept. Computing, Goldsmiths, University of London,
New Cross, London SE14 6NW, UK.
m.bishop, s.danicic@gold.ac.uk

Slawomir J. Nasuto & Richard J. Mitchell
Cybernetics Research Group, School of Systems Engineering
University of Reading, Reading, Berks, UK.
s.j.nasuto, r.j.mitchell@reading.ac.uk

May 20, 2012

Abstract

A three way collaboration between industry and two UK universities led to the development of an intelligent spend analysis system. In this paper we outline how a novel combination of the 'Decision Tree' and 'Bayesian Classifier' algorithms[1], working with 'big data' on a real world e-procurement problem, led to the development of sophisticated, A.I. driven, intelligent spend analysis software, subsequently commercially marketed throughout the UK as the 'SpendInsight' system; a system recently deployed by the UK-National Audit Office (NAO) to highlight potential savings to UK-National Health Service (NHS) procurement of £500 million per annum [14].

We subsequently investigate how deep-rooted 'institutional inertia' can often work to inhibit the full realisation of the potential economic benefits to an organisation that should accrue from the deployment of intelligent spend analysis. Finally we demonstrate how, by linking the institutional roll-out of intelligent spend analysis to an overarching green policy agenda, such barriers may be overcome. We conclude the paper by showing (perhaps counter-intuitively) that at the societal level a strong green policy agenda may realise significant benefit for both the environment and the economy.

1 Introduction

Three linked Knowledge Transfer Partnerships (KTP) between the University of Reading, Goldsmiths College and @UK PLC — a leading cloud-based electronic marketplace provider — have produced a system named SpendInsight. This system uses various Artificial Intelligence (AI) techniques to enable e-procurers to analyse their purchases and identify potentially significant savings. As an added benefit, it is also possible to estimate the carbon footprint of products and so develop an envi-

[1] Two algorithms identified in the CFP to be of specific relevance to the 2012 NICA (Nature Inspired Computing Applications) symposium.

ronmentally friendly procurement policy.

The level of automation in the spend analysis system makes it fundamentally different from competing providers, and this translates into a number of unique selling points:

- Firstly, the fast speed of the system shortens the time from analysis to report from months to days, and allows analyses to be performed more frequently. The customer is therefore better equipped to react to changing market conditions, monitor purchasing behaviour, and assess the effectiveness of procurement policies.
- Secondly, the high level of automation allows the system to produce reports in unprecedented detail. This allows the customer to drill from high level management reports right down to the original purchase order and invoice data used to generate the report. This visibility of original data provides the accountability necessary to identify genuine savings opportunities, quantify them accurately, and substantiate conclusions reached from the analysis.
- Thirdly, detailed reports empower procurement professionals to draw their own conclusions about their own data, removing any need for expensive external consultancy.

2 Background

During the three year period of development of the SpendInsight project, as a result of data processing carried out in development, opportunities arose which allowed data to be obtained about procurements in the NHS. The application for the project subsequently focused towards analysis of spend for e-procurement for the NHS.

One component of the project focused on matching companies and products identified in the spend analysis, a second component, classification, had to work on the product data so returned [11], and a third component, ranking, focused on automatic detection of attribute data in textual descriptions of products [1].

All three components worked together to form an integrated system, which has been named SpendInsight. Key to the system is the ability to classify vast numbers of different products from a variety of suppliers and hence determine equivalent products from different suppliers. Given this, it is then possible to assess the economic cost of each product and hence choose the cheapest.

The core system first went live in 2007 when the company created a repository of company and product information and, importantly, a system for identifying duplicate companies and products. Commercial opportunities for the de-duplication technology were subsequently developed which, in turn, meant that the de-duplication system became an important technology which needed to scale with the overall system, whilst maintaining traceability from input-data to output.

The scalable de-duplication technology enabled the deployment of a large-scale spend analysis solution across NHS trusts in London; this work highlighted potentially large scale savings in NHS purchasing. This result was subsequently independently affirmed in the National Audit Office report of February 2011 [14], which concludes that, rolled out across all NHS trusts in England, annual savings of £500 million pounds could be made (over 10% of NHS spending on consumables). In addition, using related 'GreenInsight' technology, the 'environmental' cost of each classified product can also be allocated, *i.e.,* GreenInsight enables e-procurers to assess the environmental cost of their purchases.

3 De-duplication

The core approach to de-duplication is to use two staging databases for the input data, at two levels of granularity. At the most detailed level are purchase orders and invoices, and each purchase order line and invoice line can be traced back to lines in clients' data files — which are typically received in CSV format.

The first challenge, then, is data integration. First, data must be extracted from diverse client systems. Although normally delivered in CSV files, the number of files, the columns in the files, and relations between the files are typically peculiar to each client. (It is the first author's experience that no two installations of the Oracle 'iProcurement' system are the same).

Once received, data must be stored in a single unified schema to allow it to be queried. However, this is insufficient to be able to generate useful reports; at this point

the duplication problem becomes apparent. Even within the finance system of a single organisation a single supplier may appear more than once — each occurrence with a subtly different name, *e.g.,* 'Limited' vs. 'Ltd.' Of course the problem is exaggerated when comparing across organisations. Furthermore, one must first successfully identify multiple instances of the same supplier before proceeding to the harder problem of identifying the unique products they sell.

3.1 Hierarchy of abstraction

When data files are received they are first loaded into a 'raw data' database. Each file is scanned and an SQL table definition statement is created for the file. The table is created and the file is loaded into it. This allows data types to be determined for each column, and to check any referential integrity constraints between the files. For example, it may turn out that a purchase order in one file gives an ID for a supplier but that ID does not exist in the supplier file.

The second step is to transform the raw data into the standard format. This can be as simple as specifying a map between the columns in the client's file and the columns supported in the system. However, more elaborate queries may need to be developed — particularly when relations between data in the input files must be used. In very rare cases it is necessary to pre-process the clients' data files — for example when data rows are interleaved with 'sub-total' rows.

The first level of abstraction models companies and the products they supply. In this model many purchase order lines may 'point' at the same product, and in turn a product 'points' at a supplier. This abstraction is key to achieving scalability. From this model the system builds a 'cleansed view' of the database which is used for driving reporting for clients. In the cleansed view the suppliers have been de-duplicated and, in turn, so have the products they supply.

The cleansed view maintains an 'audit trail' of the matching performed and the evidence upon which matching was based. This is important if clients query results because the results can always be traced back to the original data. In the final reporting, clients can see how suppliers and products have been matched, and can supply feedback to the system by identifying false-positive matches and additional matches. Because the 'cleansed view' is separate from the data itself, there is a complete 'audit trail.'

3.2 Rule engine

The cleansed view is built by a rule engine. All of the rules are applied, iteratively, until the system stabilises. Each rule may use information in both of the staging databases and in the partially built cleansed view (a kind of feedback loop), and the rule may make use of additional custom indices built on these data.

The staging databases have increased in size tenfold over the last three years, but the processing time to build the cleansed view has not increased significantly. Typically, using current technology the cleansed view can be re-built from scratch in under one week.

4 Automatic classification

The automatic classification is the other substantial component of the system. This was developed in parallel to the de-duplication technology. The core technology allows procurers to identify and cost 'equivalent' products, with an extension offering 'carbon analysis' of purchasing decisions enabling procurers to analyse both the economic and environmental cost of purchases.

Text classification is the task of predicting the class of a previously-unseen document based upon its words. The relationship between words and class is learnt from a labelled training set. Since the 1960s, many methods have been proposed, including decision rule classifiers [4], decision trees [15], k-nearest neighbour [16], Naïve Bayes [8], neural networks [13], regression models [17], Rocchio [5], the support vector machine (SVM) [7] and winno [2]. For this work the classification task is to assign each product in the cleansed view into one of about 2,000 different classes. The main data source upon which the classification task draws is the free-text descriptions on purchase order lines. In the cleansed view there may be hundreds of different descriptions for the same product.

Naïve Bayes is a probabilistic classifier that has been used since the early 1960s [10]. It has advantages over other classifiers in its simplicity, learning speed, classification speed, and storage space requirements [3]. In

the multi-variate Bernoulli event model, a document is a binary vector over the space of words [10] *i.e.,* each dimension of the space corresponds to a word. The words are assumed to be dependent only on the class to which their document belongs — an assumption which clearly is false. This is the naïve step. Nevertheless, Naïve Bayes has empirically outperformed many other algorithms [6, 9, 10, 3].

In our work the descriptions for each product are reduced to a *bag of words* and a set of manually classified *training data* is used to calculate the conditional probability of a word belonging to (a product in) each class. An application of Bayes Theorem gives the conditional probability of a class given a word; thus words and, in turn, products can be classified.

While in principle the Naïve Bayes classifier can be applied to the whole data set, in practice the applicability is limited by the training data set. If there is little or no training data for a particular class then the probability of the class being chosen is small. In one example we found 'bone granules' classified as food because the only 'granulated' product in the training data was gravy.

The training data set has been expanded considerably since the start of the project, and wherever possible assimilates clients' classified data sets. However, here we have the additional problem of deciding whether we believe that clients' data has been accurately classified. We found useful heuristics for this decision problem to include the proportion of products classified to non-existent classes, and the internal consistency of the classification, *i.e.,* to how many different classes has the same product been assigned.

Even when clients' data is not suitable for use as training data, it can still be used to provide 'additional guidance'. Two kinds of extra guidance data are used by the SpendInsight system. Both make use of the three-level hierarchy of the 2,000 classes. The first is product type: the system may deduce that a certain product should be classified into a specific level 1 or 2 class. Such deductions can be made where a product has previously been classified inconsistently between different clients or sources. This extra information is used to limit the classes into which the Naïve Bayes algorithm may assign the product. The second type of extra guidance is at supplier level: the system may either deduce [or be told] that a specific supplier either only supplies products in or does not supply products in certain level 1 classes. Such information is typically used to limit the classifications permitted by the Naïve Bayes algorithm.

The second major *NICA* algorithm employed for classification is the 'Decision Tree' which is a divide and conquer approach. The node at the root of the tree divides the training set into two subsets. In text classification, one set usually contains those training documents with a certain word, and the other contains those without. Both of these subsets are further split at the root's children, further split at the root's grandchildren, and so forth down the tree. A leaf is usually formed where the set of training documents are all of the same class, and that class is assigned to the leaf as a label.

One of the most popular decision tree algorithms is Quinlan's C4.5 [15] which uses an error-based pruning algorithm.

In our work the Decision Trees are taught from the training data by the symbolic rule induction system (SRIS). Decision Trees have been built for only 19 classes, all of which have been carefully selected (by hand) to have good supporting training data and be among the classes that are more highly visible to clients. Building and testing a decision tree can take many weeks, which limits the rate at which they can be added to the system.

Finally, for each product the classification system must decide between the various candidate classes suggested by each of the algorithms. Essentially, there is a trade-off between the accuracy of a method and the proportion of the dataset to which it can be applied. For example Naïve Bayes is the least accurate method but can be applied to the whole of the data set, whereas a Decision Trees can be much more accurate but applied only to relatively few classes. The total time taken to apply classification is dominated by the Naïve Bayes algorithm, which can classify about 100 descriptions per second.

5 Intelligent spend-analysis — barriers to full impact: changing purchasing behaviour

At the launch of the software industry standard for green data interchange (RSA London, 15/11/11) Ronald Duncan of @UK PLC reported how AI technology in classifi-

cation and matching was exploited in a new 'spend analysis' system co- developed with the University of Reading and Goldsmiths College.

By highlighting how equivalent products can be bought at the cheapest available price, the SpendInsight software was successfully deployed by the UK NAO[2] to generate the evidence base for its recent report [14] assaying huge potentially savings (of around £500 million per annum) to UK NHS procurement[3]. The fall-out from this research prompting widespread coverage in the UK news media[4] and a subsequent exposé by the BBC radio 'File on Four' programme detailing inefficiencies in NHS procurement[5], with the impact of the work subsequently meriting discussion in the UK parliament[6]. And yet, perhaps surprisingly given the current economic climate, data in the NAO report details only 61 of UK Health Trusts currently deploying the SpendInsight system.

At first sight the huge potential savings highlight by intelligent spend-analysis software might imply organisations would rush to embrace intelligent spend analysis technology, however @UK PLC experience in deploying the SpendInsight system with the NHS procurement suggest that this is not always the case; factors other than raw cost are often important in institutionalised purchasing environments.

[2]The NAO report states, *"@UK PLC uses its in-house artificial intelligence system to classify every purchase order line raised by the trust in a twelve month period to a unique product code. This system also extracts information on supplier, cost, date and quantity ordered."*

[3]*"...accounting for £4.6 billion in expenditure, using a conservative estimate of 10 per cent, savings of around £500 million could be made."*

[4]BBC News:
<http://www.bbc.co.uk/news/health-14971984>
< http://www.bbc.co.uk/news/health-12338984>.

[5]BBC Radio File on 4:
<http://downloads.bbc.co.uk/podcasts/radio4/fileon4/fileon4_20110927-2045a.mp3>.

[6]House of Commons Committee of Public Accounts, Formal Minutes Session 2010–12, "TUESDAY 15 March 2011. Members present: Mrs Margaret Hodge, in the Chair, Mr Richard Bacon, Stephen Barclay, Matthew Hancock, Jo Johnson, Mrs Anne McGuire, Austin Mitchell, Nick Smith, Ian Swales, James Wharton. Agenda item (1) NHS Trust Procurement. Amyas Morse, Comptroller and Auditor General, Gabrielle Cohen, Assistant Auditor General and Mark Davies, Director, National Audit Office were in attendance. The Comptroller and Auditor General's Report NHS Trust Procurement was considered. Sir David Nicholson KCB CBE, Chief Executive, NHS, David Flory CBE, Deputy Chief, Executive, NHS, Peter Coates CBE, Commercial Director, and Howard Rolfe, Procurement Director, gave oral evidence (HC 875-i). [Adjourned till Wednesday 16 March at 15.00pm]."

Thus, results from a recent survey[7] of attitudes to individual and organisation change, amongst 229 civil servants involved in the purchasing process, clearly showed that whilst individual buying behaviour is strongly correlated with cost, only 57% of the respondents reported cost criteria alone were enough to 'probably or definitely' change organisational purchasing decisions. Conversely, the results suggested that by linking economic savings with improved sustainability (*e.g.,* a lower carbon footprint), there was a significant increase, with 84% of respondents assessing that their organisation would 'probably or definitely' change purchasing choices[8].

6 Conclusions

The survey of purchasing behaviour reported in this paper unambiguously demonstrated that sustainability issues are a strong motivating factor in changing buying behaviour; linking economic and green spend analysis may speed up and unblock process change within an organisation. Clearly, given the large potential savings identified by *intelligent spend analysis* compared with the relatively small incremental cost (per product) of carbon off-setting, serious consideration of green issues can easily result in substantial economic benefit to an organisation. Thus, at the societal level, a strong green policy agenda may realise significant benefit for both the environment and the economy.

References

[1] Brown, M. J. Automatic production of property structure from natural language, PhD Thesis, University of Reading 2011.

[2] Dagan, I., Karov, Y. and Roth, D. Mistake-driven learning in text categorization. In Claire Cardie and Ralph Weischedel, editors, Proceedings of

[7]Survey carried out for @UK PLC at the 'Civil Service Live' conference, Olympia 7-9 July, 2011.

[8]Further support for this finding is that the result aligns well with the experience of @UK PLC in the rollout of the 'GreenInsight' [environmental analysis] software, as part of the NHS 'Carbon Footprint project'; at an organisational purchasing level, the GreenInsight software provides a fully automated environmental analysis.

EMNLP-97, 2nd Conference on Empirical Methods in Natural Language Processing, pages 55 -63, Providence, US, 1997. Association for Computational Linguistics, Morristown, US.

[3] Domingos, P. and Pazzani, M. J. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29(2–3):103–130, 1997.

[4] Furnkranz, J. and Widmer, G. Incremental reduced error pruning. In Proceedings of ICML-94, the 11th International Conference on Machine Learning, pages 70–77, New Brunswick, NJ, 1994.

[5] Ittner, D. J., Lewis, D. D. and Ahn, D. Text categorization of low quality images. In Proceedings of SDAIR-95, the 4th Annual Symposium on Document Analysis and Information Retrieval, pages 301–315, Las Vegas, US, 1995.

[6] Joachims, T. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In Douglas H. Fisher, editor, Proceedings of ICML-97, the 14th International Conference on Machine Learning, pages 143–151, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US.

[7] Joachims, T. Learning to Classify Text Using Support Vector Machines. Kluwer Academic Publishers, 2001.

[8] Lewis, D. D. Naïve (Bayes) at forty: The independence assumption in information retrieval. In Claire Nedellec and Celine Rouveirol, editors, Proceedings of ECML-98, 10th European Conference on Machine Learning, pages 4–15, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[9] Lewis, D. D. and Ringuette, M. A comparison of two learning algorithms for text categorization. In Proceedings of SDAIR-94, the 3rd Annual Symposium on Document Analysis and Information Retrieval, pages 81–93, Las Vegas, US, 1994.

[10] McCallum A. and Nigam K. A comparison of event models for naive Bayes text classification. In Proceedings of AAAI-98, the 15th National Conference on Artificial Intelligence, 1998.

[11] Roberts, P. J. Automatic Product Classification. PhD thesis, University of Reading, 2011.

[12] Roberts, P. J. et al.,Identifying problematic classes in text classification, Proceedings of 9th IEEE International Conference on Cybernetic Intelligent Systems (CIS), 2010.

[13] Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6):386–408, 1958.

[14] Report by the Comptroller and Auditor General, National Audit Office: the procurement of consumables by NHS acute and Foundation trusts, HC 705 Session 2010–2011, 2 February 2011, Department of Health, UK.

[15] Quinlan, J.R. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[16] Yang, Y. Expert network: effective and efficient learning from human decisions in text categorization and retrieval In Proceedings of SIGIR-94, the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 13–22, 1994.

[17] Yang, Y. and Chute, C.G. A linear least squares fit mapping method for information retrieval from natural language texts. In Proceedings of COLING-92, the 14th Conference on Computational Linguistics, pages 447–453, 1992.