From the Data Mine to the Knowledge Mill: Applying the Principles of Lexical Analysis to the Data Mining and Knowledge Discovery Process

Jean Moscarola and Richard Bolden

Jean Moscarola is Professor of Management and Business Administration at the University of Savoie, France; and Richard Bolden (MA) works for Le Sphinx Développement, France.

Please contact: Richard Bolden
Le Sphinx Développement, 7 rue Blaise Pascal, 74600 Seynod, France.
Tel: 04 50 69 82 98 Fax: 04 50 69 82 78
Email: rbolden@lesphinx-developpement.fr

Abstract. This paper argues that the traditional approach to datamining is dominated by quantitative tools which assume knowledge to be inherent in the data: the data miners task simply being to find it. We propose, however, that true knowledge arises from an interaction between the information and the user.

The notion of user interaction in datamining demands a modified approach. An environment must be developed in which the user is encouraged to participate in an interactive learning cycle, where knowledge is progressively extracted from the data. The combined techniques of lexical approximation, hyper-text navigation and quantitative statistics can form the foundation stones of this "knowledge mill" by permitting a progressive entry into the information and the identification of trends not readily visible via other techniques. Such practices are no longer the exclusive domain of large corporations with in-house databases, but open to anyone wishing to explore internal or external data sets.

Introduction: the "Data Mine"

The notion of datamining arises from the duel effects of needs and opportunities: the need to follow and adapt to rapid changes in science, technology, markets, etc. and the opportunity offered by the existence of increasingly advanced, accessible and well documented information systems. Nowadays, the most recent discoveries may be found on bibliographic databases and the strategies of competitors revealed by examination of patent deposits, the Internet or other sources. Likewise,

organisations with extensive in-house data sets of their own are finding them to contain important indicators of consumer behaviour and market trends. Information is becoming a strategic advantage [1,2] and hence also is the extent to which relevant and useful knowledge can be extracted from such resources [3,4].

A variety of methods and tools exist to guide the "data miner" in their quest of drawing useful knowledge from the voluminous data banks now available, but the investment, infrastructure and training often required tend to exclude smaller players from the game. Likewise, there is a tendency to focus on large-scale "number-crunching" activities in preference to the more ambiguous, yet rich domain of text analysis.

This paper presents the methods of statistical analysis applied to lexical approximation and hyper-text navigation as a solution to the general problem of making datamining accessible to the masses, and more importantly, of tapping into the knowledge which would otherwise remain buried within poorly structured text bases. Via these techniques the user is drawn into an interactive learning cycle, not only greatly facilitating the task of finding relevant information but also offering tools for the processing and extraction of knowledge not readily available by other means.

Who uses Datamining ?

Datamining has progressed a long-way since its origins in artificial intelligence (a domain largely closed to those outside the disciplines of I.T. and statistics). We now find comparatively user friendly applications which run on desktop PCs, although, where one wishes to use multiple search algorithms more advanced hardware and a great deal of external help and training are still required [5].

No matter what the system, however, the financial investment is nearly always high. The initial outlay on the soft and hardware itself is only the beginning. Consultants and training will often be required and an appropriate infrastructure established. Packages such as Darwin, for example, require at least three staff: one who creates the models, one who specifies the business rules and one who understands the database [6]. Likewise, the majority of large-scale datamining applications are designed for use on in-house data sets: sales records, client information, etc. This implies the need for additional personnel and investment in establishing and maintaining a comprehensive database. The susceptibility of datamining to "dirty data" also calls for extensive "data cleansing" operations. Aaron Zornes, an executive of the META Group, states that 60-80% of the datamining investment is usually in data preparation [6].

Such factors, thus, make it no surprise that large-scale datamining tends to be restricted to major corporations with clear productivity-oriented objectives. We hear many case studies of datamining in the banking and commercial sectors, focussing on topics such as: market segmentation, customer loyalty, fraud detection and marketing [7] but few in other sectors, such as health-care and education, where the data is available but not the money or personnel.

Extending the Domain of Datamining

We see a number of ways in which the datamining process could be extended to meet the needs of different users. Tools should:

- match the investment potential of small or non-profit orientated users;
- permit access to external data sources such as the Internet, CD-ROM and commercially and publicly available databases;
- require a minimum of data preparation, even for unstructured/deteriorated sources;
- draw on the immense knowledge contained within open-ended text;
- be easy to use and produce results which are intuitive to interpret;
- encourage user involvement such that knowledge can be constructed interactively, driven by needs rather than a computerized algorithm;
- offer a greater capacity for hypothesis formulation and testing.

Such tools would offer an intermediate level of datamining, falling somewhere between the full-scale corporate model and the heterogeneous mix of data management, processing and analysis software often promoted under the guise of "datamining".

Lexical Analysis: A Cross Between Quantitative and Qualitative Analysis

When exploring external data sets we face two distinct problems: (a) how do we find the information sources? and (b) how do we extract knowledge from them? By permitting the almost instant access to increasingly large information sources, the motors and engines of research (such as those found on the Internet) contribute strongly to solution of the first problem, but aggravate the second.

Traditional analysis is no longer viable in this context. Quantitative tools are seriously handicapped by their dependency on clean, well structured data and the inability to deal appropriately with open-ended text; whilst quantitative approaches (content analysis, trees and nodes, etc.) are far too labour intensive to be applied haphazardly to texts from relatively unknown sources.

Lexical analysis offers a middle-ground between quantitative and qualitative analysis, being rapidly applicable to texts of all types, and giving a far more flexible interface between the tasks of data acquisition, analysis and interpretation. This approach is typified by the calculation of "word lexicons": lists of words and their corresponding frequencies in the corpus. By viewing the most common or distinctive terms in a text one can rapidly determine the overall subject matter without the need for an in-depth reading.

The power of lexical analysis is greatly enhanced by the use of word dictionaries and hyper-text navigation. Dictionaries permit the suppression, highlighting, or grouping of terms; whilst hyper-text navigation offers a useful means for selective reading of text directed by the marked word lexicon.

What's more, lexical analysis can be applied to poorly maintained or unstructured data. Whilst it remains be true that if you wish to perform detailed, accurate

analysis, some data preparation is necessary; for the preliminary exploration of a data set very little work needs to be done. In this manner you can select data almost at random and use the results of your lexical analysis to decide whether or not to continue.

User Involvement in the Datamining Process

Algorithmic datamining approaches (neural networks, decision trees, clustering, fuzzy logic, etc.) use the computer's processing power to search for patterns in the data. They are not hypothesis-lead, rather they browse through large volumes of information in search of trends. This may be a good means of identifying unanticipated results, but is less useful when wishing to find the response to a pre-defined query. Likewise, these practices are only really applicable to very large data sets (preferably of gigabyte dimensions) otherwise there is the possibility of error.

Social scientists, for example, are wary of results which have not been hypothesis-driven. The concept of statistical probability implies that of every 100 analyses on unrelated data, five will come out significant (a 5% probability threshold). It requires the informed judgement of the analyst to decide an appropriate significance level and which results to accept or discard. Where results are outputted cold from the machine, interpretation may be difficult, however, where the researcher has been involved throughout, the task is much easier.

Lexical analysis brings the user into an interactive learning loop where the preliminary hypotheses of the researcher are refined in accordance to the word lexicon and selective viewing of text. From this, new search terms, classifications and divisions may be determined with which to return to the data set. Rapidly, the investigation is narrowed to the most important or interesting aspects of the data, or abandoned to enable the individual to move on to the next text to analyze.

From the Data to the Overview

The ultimate aims of textual data mining, however, are greater than simply providing a peep-hole through which we can view relevant information. We wish to move beyond the raw text to gain a global understanding, to identify major themes and to distinguish systematic developments, trends and changes from the mass of information; in brief, we hope to pass from a preoccupation with detail to the "big picture". Such a level of knowledge is becoming essential for effective strategic reflection and one of the huge "prizes" of textual data mining.

Simply reading the text is no longer sufficient at this stage for a number of reasons:

- The quantity of information to take into consideration is too large when wishing to follow developments over a long period or from a large data set.
- The most relevant knowledge is often drawn from interactions between multiple considerations, thus complicating reading and synthesis.

- The most important effects may arise from the propagation of apparently insignificant trends or from a large body of information, which may well be missed even by the most attentive of readers.

Added to those of lexical approximation, the techniques of statistical analysis of the lexicon and textual data analysis greatly increase the potential for the researcher to draw important knowledge from computerized information sources. They enable the treatment of large bodies of information which not only would otherwise be unusable due to time restrictions, but also in which a key part of the content would escape identification from standard investigation. From the scientific analysis of textual and numerical data, the statistical approach allows one to identify and highlight structures and findings unobservable via other approaches.

Typical Lexical Analysis Procedures

The principle of lexical analysis is simple: it consists of applying quantitative analysis to the graphical forms present in a text; a "graphical form" being a continuous character string containing no separating character. Studying the statistical distribution of these forms enables the production of summaries and the identification of "significant" trends.

Stage 1: Preparing data for analysis

A number of key steps are involved when preparing a text for lexical analysis:

1. *Find the records*: the first step clearly is that of accessing the data set. Such sources may include the Internet, CD-ROM, databases, etc.
2. *Prepare the corpus*: once records have been located they need to be prepared for import by the lexical analysis software. This may simply involve saving the file in ASCII text format, and inserting markers and annotations as desired.
3. *Import the data*: structured, semi-structured and unstructured data can be opened. Variable names, types and missing values will be coded automatically. The user's task is simply to monitor the process to ensure it has been performed as intended.

Stage 2: Using the lexicon

Once the data file is established the process of lexical analysis can begin. Steps depend upon the ultimate aims of the research, however, some typical ones include:

1. *Compute the word lexicon*: lexical analysis is driven by the word lexicon. This is an automatically calculated list of graphical forms present in the corpus. It may be presented in a variety of ways, but the most common and useful is in descending order of frequency.
2. *Reduce the word lexicon*: in free text many of the most frequent terms are of little interest. These may include "tool words" (those terms vital for the construction of language but which convey little meaning) and numbers/codes. Deletion of such elements (via dictionaries and automated searches) will instantly give clearer of the specific content of the text.

3. *Mark lexicon words*: marking terms manually or via dictionaries enables further manipulation of the lexicon and provides a basis for hyper-text navigation.
4. *Group equivalent terms*: the inherent ambiguity of language can be reduced by grouping equivalent terms. This may be done automatically (by word stem), via dictionaries, or manually on-screen.
5. *Hyper-text navigation*: skim through the text selectively viewing only relevant entries (those containing marked words; those matching a specified profile; etc.).
6. *Word environment*: move beyond hyper-text navigation to display words in context and calculate "relative lexicons" (those words found before or after the chosen element).
7. *Search for expressions*: whilst individual words give a certain level of understanding, nothing speaks as loudly as expressions. A search for repeated word strings permits the extension of all the previous analyses to an expressions list.
8. *Lexical statistics*: calculation of indicators such as response banality and richness offers a new way of looking at text.
9. *Generate an index*: the word or expressions list could be used to calculate an index, permitting rapid identification of relevant observations.

Stage 3: Lexical cross-analysis

Where the power of lexical analysis becomes really clear, however, is its potential to combine the domains of quantitative and qualitative analysis. Open text variables can be crossed with closed response variables and quantitative statistics applied to the outcomes. Calculation of new variables from open text permits the application of analyses not readily available. Techniques may include:

1. *Lexical table*: a table crossing lexicon words (or expressions) with value labels of a closed response variable.
2. *Specific words table*: a search for the most distinctive terms for each modality of a closed response variable.
3. *Contextual summary*: lexical indicators calculated for each modality of a closed response variable.
4. *Cross analysis*: the recoding of open text to a closed response variable permits application of standard statistical techniques (Chi^2, t-test, etc.).
5. *Data visualisation*: techniques such as "factor mapping" permit a new, graphic representation of data originating from texts.

Example: Analysis of a Bibliographic Database

Here is an example of how lexical analysis could be applied to the exploration of a bibliographic database.

Data preparation and import

The Social Research Methodology database[i] is a CD-ROM directory containing details of over 34,000 publications in the domain of the social and behavioral

sciences. A search was run to find entries classed under the category of "behavioral research" and the 278 corresponding records exported to an ASCII text file.

The text file was then opened directly by the analysis software[ii], simply by indicating which marker characters should be used (fields were preceded by a carriage return and followed by a colon). Variable names and types were identified automatically, with 18 different fields being detected (many of these differing according to whether the entry was a book or journal article). Fields available for nearly all entries were "Title", "Author", "Abstract" and "Date of publication". Missing values were recorded where fields were not present.

Variable modification

In order to permit simultaneous examination of all relevant text, fields containing similar information were combined; thus a new variable was calculated in which the fields "Title" and "Abstract" were merged. A second modification reduced the open-ended numerical variable "Date of publication" to a closed scaled variable in which dates were restricted to five comparable categories.

Generating the word lexicon

Once the data was ready the word lexicon was calculated. The list was generated automatically and reduced to eliminate "tool words", words containing a number and words of fewer than three letters.

Prior to lexicon reduction the 10 most frequent words were as follows: of (682), the (525), and (495), in (297), research (293), on (232), a (190), to (168), behavioral (161), for (152). The corpus consisted of 10137 words and the lexicon of 2070 different words.

After lexicon reduction the most frequent terms became: research (293), behavioral (161), analysis (126), behavior (78), data (71), methods (69), social (50), test (44) models (43), theory (40). The valid corpus was reduced by 40% (to 6383) simply by ignoring the 100 most ambiguous terms (reduced lexicon = 1915 words). From the reduced lexicon, key issues are easy to identify: "behavior", "research", "analysis", "methods", etc.

Identifying repeated expressions

Whilst isolated words give a certain insight into the meaning of a text, expressions can give a greater "feel" for the data. An automatic search for repeated segments of text was performed, searching for strings with between 2-10 words; a minimum frequency of three and ignoring "tool words". This identified 169 distinct expressions, the most common of which included: "behavioral research", "research methods", "data analysis" and "behavioral sciences". By generating a new variable in which terms from expressions were linked, we permitted the analysis of a combined word and expressions lexicon.

Looking for specific words over time

By performing a lexical cross analysis of the words and expressions in the new text variable with the recoded date variable we were able to identify trends over time. A Specific words table listed terms in order of decreasing distinctiveness over time.

Incorporating the 10 most distinctive words per category into a cross analysis permitted generation of a "factor map", a visual presentation of the key terms by date (Fig. 1).

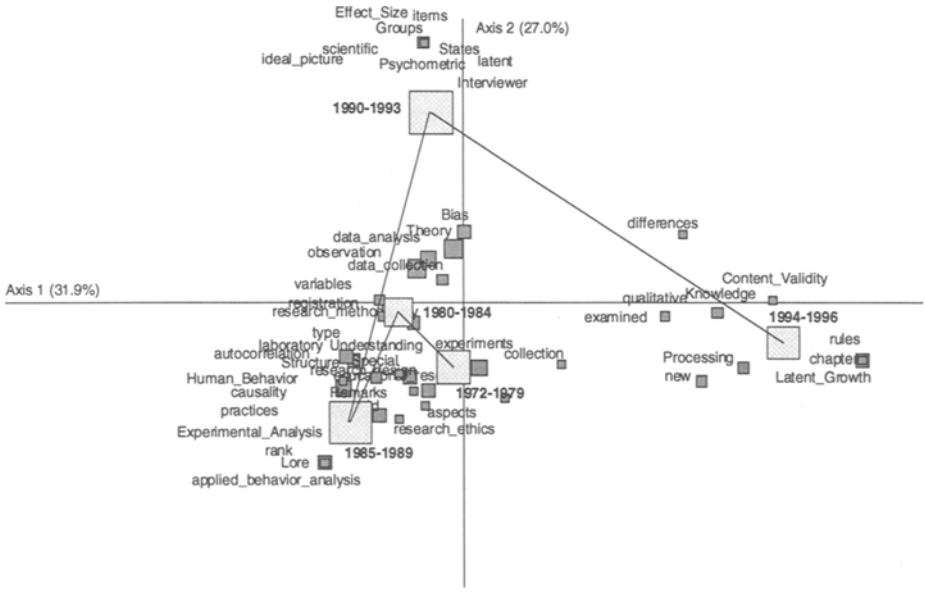

Fig. 1. Factor map of specific words by publication date.

Factor maps are intuitive to interpret: the proximity of points corresponding to the strength of relationship. We can thus conclude that throughout the 1970s and 1980s the field of behavioral research remained relatively unchanged. There was a primary focus upon experimental, quantitative approaches, combined with theoretical considerations regarding ethics, causality and research methodology. In contrast, however, throughout the 1990s there have been rapid changes. The early part of the decade (1990-1993) saw a move to consideration of psychometrics, groups, interviews, effect size, etc.; whilst the mid-part (1994-1996) brought out issues regarding qualitative analysis, content validity and latent growth.

Summary

By viewing findings in this context, the full power of lexical analysis becomes evident. We have moved from the raw text to a totally new environment in which

statistical trends, not readily available in the raw data, come to light: from the "data mine" to the "knowledge mill"!

Conclusion: the "Knowledge Mill"

The task of extracting knowledge from the datamining process requires one to address the more general question of how to derive meaningful information from numerical, textual, coded, or any other form of data. Until now there has been a tendency to restrict the analysis of single data units to the most "natural solution" - that defined by data nature. Thus, numeric data is analyzed with statistics, and texts with content analysis, reading or the application of dictionaries and thesauruses.

Such an approach may be effective for homogenous or well structured data sets, but hits a bottleneck when asked to consider poorly structured or textual information. For investigation of such sources a more interactive approach is necessary. The techniques of lexical approximation, hyper-text navigation and lexical cross-analysis offer a means for encouraging user involvement in the datamining process. In addition, they permit exploration of semi and unstructured data sets which are not suitable for investigation via standard datamining and analysis procedures.

The intuitive and flexible manner in which these procedures may be used make them ideal tools for organisations without the investment potential to install large-scale data management and datamining systems. The increased user involvement encourages a more hypothesis-lead approach to investigations and the construction of knowledge which would escape us via other approaches. In short they encourage a move from "datamining" to "knowledge milling".

References

1. Jakobiak, F. Exemples commentés de veille technologique. Organisation (1991).
2. Rouach D. La veille technologique. Puf (1996).
3. Lesca, H. and Belkatir, M. Pertinence: un instrument pour évaluer le besoin de veille stratégie de l'entreprise. Eyrolles (1991).
4. Dou H. Veille technologique et compétitivité. Dunod (1995).
5. Darling, C. Datamining for the masses. Datamation (1998).
6. Freeman, E. Datamining unearths dollars from data. Datamation (1998).
7. McCarthy, V. Strike it rich! Datamation (1998).

---

[i] The SRM database is distributed by Scolari, Sage Publications Software.

[ii] The analysis software used in this example was the *SphinxSurvey (Lexica Edition)* package, developed by Le Sphinx Développement and distributed by Scolari, Sage Publications Software.