RESEARCH PAPERS OF THE WROCŁAW UNIVERSITY OF ECONOMICS
No. 35 — 2008
Advanced Information Technologies for Management – AITM 2008

Richard F. Bonner[(1)], Tetyana I. Mamchych[(2)], Iryna I. Malchuk[(2)]

[(1)]Mälardalen University, Västerås, Sweden; [(2)]Volyn National University, Lutsk, Ukraine
richard.bonner@mdh.se; mamchych@univer.lutsk.ua; mal_i@mail.ru

ON THE PROBLEM OF MINING THE WEB – FOR A CURRICULUM[1]

Abstract: We report a pilot project, which is to examine the plausibility of automated response of a professional curriculum to the professional market by adaptive mining of the Internet. In particular, the project shall prototype software for curricula in applied statistics, founded in computational learning theory. Its technical progress is reported elsewhere; here, we put it into a bigger picture of general curriculum considerations. We outline its background, sketch its learning domain, state its ultimate criterion for self-evaluation, share its theoretical framing, and observe some principal limitations of its interest to teaching.

1. Background

Universities across Europe, recall, have in recent years been living to happy grunts of a singularly hungry organism, set to digest a continent of education, a millennium of varied custom and academic tradition, down to something of a formulary. Many seek mental shelter in the thereby supposedly arising "opportunity to re-examine curricula", bravely believing that examine and cut-and-paste appears to sound better than just cut-and-paste. In tune with these events, which is not to say in support, the authors found themselves *ab ovo* confronting questions of professional curricula [Bonner 1995; Mamchych 1997a;1997b; 2001; 2006], but this time in a principal way.

We were once again quite surprised to find so little legible to support educators in their thinking about their teaching, no theory of the curriculum that is, though the artifact is centerfold for organized education, and so crowded a topic of corridor debate. We found no words for the constituents and relationships of structures and processes, which would formally frame it. Legible descriptions of instances as in

[1] Supported by The Swedish Institute, grant SI-01424/2007.

computer science or statistics help little as they fall back on the terminology of their subjects and on their pragmatic *status quo*. The very word *curriculum* is dubious; for some it is a list of contents, for others it is an algorithm, for yet others it is the physical process of a course of study. There is no good language to talk about university teaching in abstraction of its factual context.

We do not wish to suggest that talking about teaching in abstraction of subject matter should be necessary or desirable: there is normally enough to talk about within subjects. How much management, for example, lies behind a typical curriculum in "management"? How much analysis behind one in "system analysis"? How much statistics behind one in statistics? Each could surely be improved in its proper context, would someone burn to do so. But how to "re-examine" a curriculum without abstract language if "inter-disciplinary", say if in an eclectic field of business computing?

Talking about such things is no simple matter, not even with due language, given how hard it is to talk about knowledge or technology, or the role of the university in a society. Those who do without, hamper the spirit of subjects, the culture and the dynamics, juggling the superficial, the superfluous, and the spurious. The giga-dollar flood of (re-examined?) inter-disciplinary degree propositions gushing out of the "entrepreneurial university" consists in bulk of cheap cut-and-paste jobs.

Neither did we find much legible on the conditions, actors and factors, controlling the path between a description of studies and a live graduate. Indeed, even the endpoints of the path are rarely stated, and, it appears, for good reasons. Any description of studies, which gives justice to its subject and thus makes sense to the expert, runs the risk of making sense to the expert alone. Descriptions of graduates appear as problematic. By whose preference, for example, are graduates to be ranked? Are the qualities that describe graduates to hold at the time of their graduation or later in life, and if so, when?

Now, a working educator whose job it is to study and teach a ground subject of mathematics or statistics, may perhaps find time to examine own teaching, and perhaps even to then share findings with close colleagues, but certainly not to plunge into the philosophy of education. For such an educator, the engagement in any basic discussion of curricula must end upon realizing the risks in doing it wrongly and the extent of work entailed by trying to do it right. For us it ends here.

It is another matter altogether for us to raise a pragmatic question that is well defined and that fits our schooling. This we do now. Towards the end of our discussion, we point to some principal limitations for this line of work.

All references that can readily be found with Google have been skipped.

2. The responsive educator

Consider an institutional educator who rejects any civil responsibility over and above the provision of employable graduates. For such an educator, the curriculum

problems simplify dramatically and may indeed be talked about in some generality. For, disregarding as we now must the personal, the political, and all the other extraneous unavoidably present factors, an employable graduate is but a graduate who appears to some professional employer apt to perform the duties of the employer's profession.

Assuming that the words the employer and the applicant utter at the interview as well as the actions they may each perform would agree in connotation if they agree as sounds and motions, it then follows that the ideal distribution of themes encountered by the student in the teaching towards a degree in such and such a profession should simply be the same as the statistical distribution of the themes encountered by the entry-level professional in the practice of the same profession.

Call a distribution of the former kind a "curriculum", and a distribution of the latter kind, a "market". The task of the responsive educator in proposing a curriculum is then but the orthodox task of sampling the market. A market may here be set by quite arbitrary criteria, which makes it slightly wider a concept than the traditional; it could, for example, be defined by the distribution of mobile telecommunication techniques used by dentists in Lutsk. Though formally well defined, in reality such a market may not exist in any stable sense; and, if it does, it need not be of interest or subject to sampling.

In the past, the task of sampling a professional market necessarily involved extensive questioning and interviewing, employing time and effort of many people, and it was rarely undertaken by less than a government body or a professional association; the extent of work in such a task is well illustrated by the impressively explicit "Hudson Report" [Hudson 1992] of fifteen years ago, about the computing profession in Australia. Today, however, with the Internet becoming the dominant medium for professional communication and a dominant environment for professional work, and with the advent of tools for adaptive statistical analyses of massive data, it has become realistic to try to "mine" the Web in bounded time for the market. This, combined with the expectation of continued growth of the "cyber-infrastructure" [Aisworth et al. 2005], has made it plausible for an educator to consider sampling the market of any well-defined profession and updating the curriculum continuously.

One may for simplicity picture a responsive educator as a computer program that continuously learns the market. It then randomly generates exercises, questions, examples, tests, etc., for the students following the market. Superimposed on the problem of learning the market are learning problems of students, which the program controls by testing the students and individually adapting the teaching material. To write and maintain such a program employs known machine learning techniques. Such a task is quite well defined and it can be done gradually, new versions building on earlier ones, satisfactorily tested.

We, the second and third of the authors, are indeed presently engaged in early stages of writing a responsive educator for teaching statistics to prospective professionals. Applied statistics is today taught to students of many specialties. All

students learn base material, and there is broad consensus as to what this material should contain: the table of contents varies little from one textbook to another. The students then learn in greater depth what they will use more often. The economists, for example, appear to use regression techniques more often than non-parametric methods, while the opposite appears true for psychologists.

We are at the stage of free experimentation [Malchuk,Mamchych 2007]. We estimated the weight of each theme in a basic statistics course by an estimate of the relative frequency of the occurrence of the theme in the Internet. Sampling the Internet was done by hand. Using the standard platform Moodle, we then experimented with computer-generated tests, the questions sampled randomly from the distribution defined on the themes by the weights.

To put a small project into big perspective, we note with keen interest a national German project New Statistics, which "… aims at establishing a new, multimedia-based form of instruction in statistics in German universities. ... the project attempts to transform the traditionally formal, mathematically dominated statistical instruction into a problem-solving, reality-oriented approach." We here take "reality-oriented" in spirit synonymous to "responsive". We also note visible presence of commercial providers of statistical education whose business it is, as we understand it, to be responsive.

3. The domain

In hindsight, we chose the field of statistics for five reasons. First, the tasks of statistical practice are often determined by the statistical tools they use. These tools are commercially packaged as statistical software such as SAS or SPSS, or as statistical toolboxes of generic mathematical software such as Mathematica or MatLab. They are also available non-commercially from the Internet, for example as the S or R software, under the GNU initiative. In all cases, it is plausible to suppose that the extent of their use would leave traces in the Internet. And, to recall the argument, knowing which tools are used and how often, the educator can design a curriculum that duly prepares students to use them.

Our second reason theorizes the first: statistics and computation are related intimately to the point that statistical inference appears explainable by the computational principle of least description length [Cilibrasi 2007]. Third, curricula in applied statistics have been responding rather poorly to the dramatic changes of the recent years in the market. Fourth, considering the kinship of statistics and learning [Nakhaeizadeh, Taylor 1997], why not use statistical methods of learning theory to find out how statistical methods were used? Finally, teaching statistics is a main engagement of the second author.

From Wiki: "The word statistics ultimately derives from the modern Latin term *statisticum collegium* ('council of state') and the Italian word *statista* ('statesman' or 'politician'). The German *Statistik*, first introduced by Gottfried Achenwall (1749),

originally designated the analysis of data about the state. It acquired the meaning of the collection and classification of data generally in the early nineteenth century." With the advent of universal computation about three decades ago, statistics began transforming into a new field of science tied by the mathematics of computation and information.

The omnipresence of statistics in modern times is hardly surprising. The feedback loop, from observation (data), through computation (inference), to observation (new data), is the navel of science, intelligence, learning, control, management, etc. Computing also plays supportive roles: filing and searching documents, visually displaying data, and generally maintaining a mathematical and orderly environment. Statistics controls the loop: it first chooses and runs a program to take in data and output a theory, and it then tests the theory on new data and/or meta-data; if the theory passes, it stops, if not, it loops. It thereby designs experiment and interprets its multiple outcomes.

Policy papers and encyclopedias define statistics compatibly. The paper [Kettenering et al. 2004], for example, assesses the state of statistics, and we read on pages 2-3: "Statistics is the discipline concerned with the study of variability, with the study of uncertainty and with the study of decision-making in the face of uncertainty. ... A distinguishing feature of the statistics profession, and the methodology it develops, is the focus on a set of cautious principles for drawing scientific conclusions from data. This principled approach distinguishes statistics from a larger venue of data manipulation, organization, and analysis. ... When used appropriately, these tools help to curb false conclusion from data."

Quoting Wiki again: "Statistics is the science and practice of developing knowledge through the use of empirical data expressed in quantitative form. It is based on statistical theory, which is a branch of applied mathematics. Within statistical theory, randomness and uncertainty are modeled by probability theory. Because one aim of statistics is to produce the 'best' information from available data, some authors consider statistics a branch of decision theory." The *Glossary of Meteorology* of the American Meteorological Society explains wider: "The systematic analysis of random phenomena. Primarily, it is the application of probability theory to specific data, but includes special techniques and principles not subsumed under probability. Statistics is concerned with collecting and processing data, summarizing information, estimating descriptive constants (parameters), discovering empirical laws, testing hypotheses, and designing experiments in such a way that valid inferences can be drawn from empirical evidence."

Most reliably but least concisely, statistics is defined by its scientific output, indexed by the *Current Index to Statistics* (CIS), a joint venture of the American Statistical Association and the Institute of Mathematical Statistics. Quoting the CIS page: "The CIS is a bibliographic index to publications in statistics and related fields. References are drawn from 162 core journals, as of 2003, that are fully indexed, non-core journals from which articles are selected that have statistical content,

proceedings and edited books, and other sources." Smaller yet representative are proceedings of statistical conferences and annual meetings such as the EMS, together with Wiley's grand encyclopedia [Kotz et al. 2005].

At the lower end of the scale one finds field reviews, survey papers, policy-type journals, textbooks, and monographs. The influential "Odom report" [Odom 1998], prepared for the (American) National Science Foundation (NSF), assesses the state of statistics of the late 1990's among other mathematical sciences in the USA. The "Lindsay report" [Kettenering et al. 2004], in shorter version [Lindsay et al. 2004], is an outgrowth of a strategic workshop arranged in 2002 by the NSF. The collection of papers [Raftery 2001] surveys the directions in which statistical science develops. Special journals, such as Springer's *CHANCE* and the *Statistical Journal of the United Nations Economic Commission for Europe*, monitor border areas. Recent book titles in orthodox statistics, note, are few as compared to those in statistics developed in other fields.

It is the software that ultimately defines a computational science: programming languages, packages, working environments, benchmark data, etc. The major commercial packages in statistics are well known, and prime statistical software is free to download from the Web. Notable is the GNU project R that extends the S language; note also the Omega Project for Statistical Computing. Note, however, that generic statistical software is grossly overshadowed by swarms of special-purpose packages.

4. Theoretical framing

Technical work on a program termed "responsive educator" for the domain of statistics, reported in [Malchuk, Mamchych 2007], progresses along with theoretical considerations. Two groups of questions appear. The first group formally belongs to learning theory. The second group belongs to theory of knowledge, its representation, and, necessarily, its purpose.

The questions in the first group concern in the first instance the encoding of the domain of learning and the formal statement of learning problems. We have not yet confronted any questions of statistical learnability of objects, that is, questions about the initial indeterminacy of objects to be learned as measured by the learning time. Complexity issues that must arise in full-blown models do not arise in prototypes controlled by hand. Nevertheless, it is good for general orientation to work in the learning-theoretical framework from the outset.

The full domain of statistics cannot be formally encoded, but neither is it our wish to encode all. It is enough for a start to encode the domain of a definite course. For a basic course, for example, the domain may initially be defined as a formal concept structure in the sense of [Ganter, Wille 1999], following the table of contents of a basic textbook such as [Wasserman 2004]. As ground classes one may take problems and tools, in the usual sense of these words in statistical practice, both entities

being composite. Problems come in guises of varying size and complexity, from a teaching "example" to a full-blown "application" with real data. The problems and the tools may then be grouped, individually or in combination, into tasks and toolboxes, which may in turn be grouped into themes, chapters, application areas, etc. It is not essential that groupings are optimal, but they should cover all notions, and their relative links should be clear.

The non-uniqueness of vocabulary is a complication to be addressed early, as essentially the same tools or problems occur in different guises in different fields, and one word may sometimes stand for more than one thing. Initially, a human expert may set up a dictionary informally, to be adaptively modified by learning. It is essential not to forget at what level of universality our system is to operate, as our goal is double: is out system to improve a concrete curriculum, or is it to address the general (educator) problem? Toy versions of the latter can be played with by hand and they should be; full-size versions must be addressed by due method in due time. The general case must furthermore build on epistemological theory, with knowledge represented formally by logical structure.

With the domain formally structured, one may consider the dynamics. Modulo the structure, this is usual sampling theory and statistical learning. For general concept structure there is no learning theory though learning in graphs is not new. A general analysis of learning time in a concept structure appears at the moment quite out of reach.

The questions in the second group start with the naïve. How often is a tool used, and in what way? What value is there in using it? How well must one know a tool in order to use it? Can it be misused? Can another tool be used instead? Initially informal guidance, such questions ultimately define formally our objects of interest that make up a curriculum.

Attention must be paid to side conditions always implicitly present when things concern humans rather than machines. Such are the normative requirements for curricula for purpose of professional accreditation, whether *de jure* or *de facto*. We duly note *Curriculum Guidelines* of the American Statistical Association, observing there, however, a large measure of vagueness. Some such requirements cannot be fully formal but they nevertheless take form when enforced by the professional core or by the educators; they are visible in journals such as the *Journal of Statistics Education*.

The vagueness in curriculum description is indeed convenient for the human educator, whether personal or institutional. It is in practice impossible to follow tight rules, and vagueness helps avoid legal and moral consequences of non-compliance. Yet, the vague may be useful to the wise. Consider the non-controversial maxim: the graduate should know what he knows, what is known, and what could be known. Or: the graduate should know roughly how everything works, to the extent to be able, when necessary, to effectively find out from available sources how anything works exactly. Or: "(statistics) is about doing the right thing when interpreting em-

pirical information" [Vardeman, Morris 2003, p. 21]. But try explaining these to a machine!

The electronic educator, now normatively subdued and responsive to market, must also respect human limitations which differ from those of machines. Unlike machines, humans make sense of a complex problem only by viewing its simple cases. The relationship of continuous and discrete models in science may thus appear puzzling, the former dominating classical physics, the latter – modern computing, until recalled that combinatorial models, often trivial for machines to work with, may be quite impossible for people to follow.

It is obviously not realistic for the electronic educator to learn the "human factors" from sampling low-level data. It must either be taught these by a human educator, or it must be conditioned to learn them adaptively from students. We do not know yet how best to interface this expert-system type of learning with the statistical learning from samples.

5. Prospects

It is too early to be optimistic about the future of responsive education in the sense considered here, where the sole true actor is the educator, while the students are passive and the job market is uncooperative, the former being served a ready basket of curricula arrived at *via* investigative sampling of the latter. It is easier to be optimistic about educators' responsiveness to a cooperative job market that puts at the educators' disposal the data it may have and thus helps determine curricula.

It is not clear however to what extent the job market could publicly cooperate with educators given that it is not cooperation but competition, or at least it is so believed, that drives it. It is then not in the interest of an employer to disclose the competence sought. Unsolicited responsiveness of an educator may qualify as industry espionage.

This conflict of interests was traditionally resolved by non-public educational arrangements, the public education stopping short of technical specialization to be later provided discretely in-house. The Swedish L M Eriksson, for example, used to house vast educational resources. Graduates of public universities were selected for employment by their marks in basic subjects such as mathematics and the natural languages, thus choosing the best prepared and most promising for their pending in-house training.

As the industrial, the professional, and the organizational know-how, as well as the means of social control, gradually transfer from people to computers, and standardization of the production and services industries dramatically reduces the demand for personal judgment, the employers' dependence on highly skilled personnel reduces likewise. The competitive know-how of enterprises now dwelling in computer software and social relations, the competence sought from university graduates by employers no longer constitutes qualified material.

Consequently, the criteria for selecting employees essentially reduce to two: the dexterity in operating computers in a domain of application, and the so-called social skills. A statistics curriculum may only address the former, as it often indeed does. It is then not clear whether any fine-tuning of the curriculum would matter at all for the graduates' employment prospects. If it does matter, however, there is hope that the employers, when prompted, may provide open access to appropriate data that could be mined.

6. Principal limitations

Contrasting the responsive educator is the principal one, who evaluates the teaching less by employment statistics of its graduates than by academic principles. No educator, we wish to believe, may remain fully responsive or fully principal. The difference is, however, principal and there are consequences of employing responsive tools in principal schooling, or *vice versa*. In particular, the strong focus on computational tools in statistics, which the responsive schooling must maintain, has the principal drawback that it is all too easy to run software without thinking. It is however hard to argue principal one's standpoint that a data processing subject should not freely rely on computational tools in its practice.

One cannot here help recalling the six "general requirements for completing the curriculum" set forth by the (professional and responsive!) ACM 2005 guidelines for the subject of computer science, which the responsive educator of statistics might like to consider. First, second, and third, the guidelines expect from graduates the understanding of the scientific method, mathematical rigor, and familiarity with application. All three are obviously quite critical for statisticians. Fourth, fifth, and sixth, they expect good communication skills, and the ability to work in team and to organize own work. Could any statistician drop these?

One cannot help noting how universally desirable these qualities appear. Could any educator in any technical subject renounce any of them? One may wonder then whether it is any training or none that brings about the noble trait. Not one of these qualities appears in conscious form in a working bee, the technical perfection of which is beyond doubt, and which in eternally set conditions needs no feeling for the collective, no freedom of choice, no personal sense of purpose. The sail that the responsive educators try hoist is thus worked by two strings, of diligence and freedom, of which they pull but one.

References

Aisworth S. et al. (2005), *Cyberinfrastructure for Education and Learning for the Future: A Vision and Research Agenda*, Computing Research Association (CRA).

Bonner R. F. (1995), Mathematics for Information Systems: The MIST Project, [in:] *Proc. Australasian Conf. on Information Systems (ACIS'95), 26-29 September 1995*, Curtin University, Perth, Australia, pp. 543-565.

Bonner R.F., Mamchych T.I. (2005), Making sense of knowledge management: rational choice, learning, and their interplay, [in:] *Acquisition and Management of Knowledge*, M. Nycz, M. Owoc (Eds), Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1064, AE, Wrocław, pp. 353-366.

Cilibrasi R. (2007), *Statistical Inference through Data Compression*, PhD Dissertation, ILLC Dissertation Series DS-2007-01, Amsterdam.

Ganter B., Wille R. (1999), *Formal Concept Analysis*, Springer, Berlin.

Hudson H. (1992), *Report of the Discipline Review of Computing Studies and Information Sciences Education*, Department of Employment, Education and Training, Canberra.

Kettenering J. et al. (2004), *Statistics: Challenges and Opportunities for the Twenty-First Century. Report*, National Science Foundation Workshop, May 2002.

Kotz S. et al. (2005), *Encyclopedia of Statistical Sciences*, 2nd ed., 16 volumes, Wiley.

Lindsay B.G. et al. (2004), A Report on the Future of Statistics, *Statistical Science*, Vol. 19, No. 3, pp. 387-413.

Malchuk I., Mamchych T.I. (2007), On teaching applied statistics: differentiating the curriculum, [in:] *Scientific Bulletin of Chełm, Section of Mathematics and Computer Science*, No. 3.

Mamchych T.I. (1997a), Stochastic algorithms in programming courses, *Scientific Bulletin of the Volyn State University*, Lutsk, Vol. 4, pp. 101-102 [in Ukrainian].

Mamchych T.I. (1997b), Problems of probability theory in informatics courses, *Pedagogical Search*, Lutsk, Vol. 2, No. 14, pp. 63-64 [in Ukrainian].

Mamchych T.I. (2001), Some current problems in teaching various methods for processing statistical data, *Problems of Pedagogical Technologies, Transactions*, Lutsk, Vol. 4, pp. 92-96 [in Ukrainian].

Mamchych T.I. (2006), Mathematics for social science students: whom to teach what, how and by whom?, *Scientific Bulletin of Chełm, Section of Mathematics and Computer Science*, No. 2, pp. 97-104.

Nakhaeizadeh G., Taylor C. (1997), *Machine Learning and Statistics: The Interface*, Wiley, New York.

Odom (1998), *The "Odom Report": Report of the Senior Assessment Panel of the International Assessment of the U.S. Mathematical Sciences*, National Science Foundation, Arlington, Virginia.

Raftery A.E. (Ed.) (2001), *Statistics in the 21st Century*, Chapman & Hall.

Vapnik V.N. (1998), *Statistical Learning Theory*, Wiley, New York.

Vardeman S.B., Morris M.D. (2003), Statistics and ethics: Some advice for young statisticians, *The American Statistician*, Vol. 57, No. 1, pp. 21-26.

Wasserman L. (2004), *All of Statistics*, Springer, Berlin.