Pergamon

Technovation 21 (2001) 689–693

technovation

www.elsevier.com/locate/technovation

Predicting emerging technologies with the aid of text-based data mining: the micro approach

N.R. Smalheiser *

Department of Psychiatry, University of Illinois at Chicago, MC912, 1601 W. Taylor Street, Chicago, IL 60612, USA

Accepted 11 April 2001

Abstract

Text data mining should be useful for anticipating new technologies and new uses for existing technologies, insofar as one can attempt to connect complementary pieces of information across two different domains, or subsets, of the scientific literature. The present study attempted to predict genetic engineering technologies that may impact on viral warfare in the future. The analysis was carried out using a combination of conventional Medline searches and the package of advanced informatics techniques known collectively as Arrowsmith. The findings strongly indicate that genetic packaging technologies such as DEAE-dextran, cationic liposomes and cyclodextrins are plausible candidates to enhance infections caused by viruses delivered via an aerosol route — despite the fact that no studies have yet been reported that have examined this issue directly, and certainly not in the contexts of viral disease or viral warfare. The critical factor was the overall strategy of approaching the problem: first, to define two specific fields explicitly (in this case, genetic engineering and viral warfare) that are hypothesized to contain complementary information; second, to identify common factors that bridge the two disciplines (i.e. research on viruses); and third, to progressively shape the query once initial findings are obtained. Thus, in contrast to some current perceptions, the process of text data mining is neither automatic nor is it restricted to those who have access to macro analyses using customized computer systems.

Keywords: Informatics; Information retrieval; Arrowsmith; Forecasting; Prediction

1. Introduction

Technological innovation often proceeds by applying advances made in one field to a separate arena. Once the innovation is implemented, the transfer of knowledge may appear obvious or even inevitable, but without the benefit of hindsight it is surprisingly difficult to identify specific technologies that are ripe for transfer. One must simultaneously identify a need in one domain and a tool in another, possibly quite disparate domain that potentially satisfies that need — and such a task requires more than expert knowledge. A large body of research knowledge is published in the form of papers and technical reports that are accessible via bibliographic databases, leading several workers to advocate the development of techniques for knowledge discovery in databases (Fayyad and Uthurusamy, 1999), and in particular, strategies for text data mining (Swanson and Smalheiser, 1997; Hearst, 1999; Kostoff, 1999), in order to 'discover' useful knowledge that is implicit within the published record. Anticipating new technologies and new uses for existing technologies should be ideal applications for text data mining, insofar as one can attempt to connect complementary pieces of information across two different domains, or subsets, of the scientific literature, that may not have been noticed by workers beforehand.

Text data mining strategies can be divided into two types, macro and micro. Macro analyses perform data-crunching operations over a large, often global set of papers encompassing one or more fields, in order to identify large-scale trends or to classify and organize the literature. Several examples of macro analyses have been published by ourselves and others (Swanson and Smalheiser, 1997; Kostoff, 1999). In contrast, micro analyses pose a sharply focused question, in which one

* Tel.: +1-312-413-4581; fax: +1-312-413-4569.
E-mail address: smalheiser@psych.uic.edu (N.R. Smalheiser).

0166-4972/01/$ - see front matter
PII: S0166-4972(01)00048-7

searches for complementary information that links two small, pre-specified fields of inquiry. We have previously shown the value of this micro approach in helping to formulate and assess hypotheses arising in biomedical research (Smalheiser and Swanson, 1998a,b), and in the present paper, it is demonstrated how the micro approach can be employed for helping make policy decisions regarding technical innovation.

Genetic engineering technologies have the capability to alter the make-up of biological organisms and thus have the potential to impact on the way that nations may conduct, and hopefully may defend against, the threat of biological warfare (BW). To anticipate possible threats that may be developing, one needs to learn what relevant genetic research is being done around the world — not only research that is explicitly intended for military applications, but also research being conducted in medical, biotechnological, public health, agricultural or zoological contexts that might be potentially applied to BW applications in the future. This is a task for military intelligence, but it can be difficult to distinguish research intended for BW from that directed toward, for example, vaccine development or gene therapy, and intelligence officers need to prioritize which kinds of genetic research are most in need of being tracked. The task is made even more difficult by the multiplicity of BW scenarios that must be considered — for example, whereas battlefield deployment of BW agents would necessarily induce acute, fulminant disease that incapacitates troops, a terrorist threat might well involve dissemination of agents that induce chronic rather than acute symptoms.

Taking the complementary approach of using informatics to predict emerging genetic engineering technologies that may impact on BW, specifically viral warfare, the question is "Given the state of published research right now, what BW applications are possible?" whether or not there is evidence that anyone is actually exploring those avenues. The analysis was carried out using a combination of conventional Medline searches and the package of advanced informatics techniques known collectively as Arrowsmith (Swanson and Smalheiser, 1997), which seek to find meaningful relationships between two largely disparate literatures or fields of inquiry — in this case, genetic engineering vs. viral warfare. Furthermore, the focus was on research findings that were so strong and consistent that they were reflected directly in the titles of papers, although the abstracts and text of key papers were also assessed when relevant.

2. Results

The first step in this analysis was to define the problem more precisely and narrowly, in order to define two subsets of the literatures on viral warfare and genetic engineering that are likely to be implicitly related to each other in a complementary fashion. Because aerosol dispersion of viruses is a major scenario for viral warfare, and because viruses must remain stable in aerosols for a significant period of time if they are to be regarded as warfare threats, it was decided to focus initially on studies of the aerosol stability of viruses. Furthermore, the entire field of genetic engineering was not examined, but initial attention was restricted to studies that had examined virulence of viruses using genetic techniques. Thus, this approach was intended to identify ways of making viruses already known to be virulent even more effective in aerosol attacks, by altering their genetic make-up so as to increase their aerosol stability. However, the information may apply to other scenarios as well, for example, making non-virulent viruses more virulent; using viruses as vectors to carry exogenous genes that encode toxins; or dispersing viral nucleic acids, rather than intact viruses, by an aerosol route as a way of infecting hosts.

The second step in the analysis was to retrieve the existing literature on ways to alter the aerosol stability of viruses. A conventional Medline search using the public PubMed search engine (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=PubMed; August 2000) posed the query "(aerosol or aerosols) AND (stability or viability or survival) AND (virus or viral)", resulting in ~200 records that were inspected manually. As an alternative method of identifying relevant papers on this topic, an Arrowsmith search was carried out where one literature A=papers dealing with aerosol stability/survival of viruses and a second literature C=genetics/genetic techniques and virulence of viruses; title words that were shared in both literatures were denoted as B-terms and we examined the records whose titles contained B-terms that corresponded to the names of individual viruses (see details in Swanson and Smalheiser, 1997). Both approaches gave similar results: Previous research has shown that the aerosol stability of viruses can be altered by varying the relative humidity and the temperature of the air, as well as by adding compounds such as DMSO, inositol and polyethylene glycol which retain adherent water and prevent dehydration of the virus with subsequent denaturation of viral proteins (a critical factor in loss of virulence in aerosols). No indication was evident from the current literature that manipulating the genetic make-up of intact viruses could affect their aerosol stability. However, a series of papers was found showing that the aerosol stability of naked nucleic acids including viral DNA and RNA can be enhanced using packaging technologies such as the use of DEAE-dextran or cationic liposomes, that are used extensively for gene transfer experiments.

This led us to revise the query, to assess whether one or more of these packaging technologies might plausibly