A process for mining science & technology documents databases, illustrated for the case of “knowledge discovery and data mining”

Donghua Zhu
Alan Porter
Scott Cunningham
Judith Carlisie
Anustup Nayak

Abstract

This paper presents a process of mining research & development abstract databases to profile current status and to project potential developments for target technologies, The process is called “technology opportunities analysis.” This article steps through the process using a sample data set of abstracts from the INSPEC database on the topic o “knowledge discovery and data mining.” The paper offers a set of specific indicators suitable for mining such databases to understand innovation prospects. In illustrating the uses of such indicators, it offers some insights into the status of knowledge discovery research.*

INTRODUCTION

Knowledge discovery can be applied to the mining of science and technology databases (e.g., publication or patent abstract databases such as *Engineering Index, MEDLINE, or U.S. Patents).* Such analyses are known as “bibliometrics” or “scientometrics.” Results can be of value to researchers, research managers, technology information specialists (e.g., those licensing technologies), and strategic planners. These professionals must keep abreast of rapidly changing technical domains and anticipate future developments. This article introduces a process of knowledge discovery that we call “technology opportunities analysis” (“TOA”). It shows some indicators (operational measures of performance) generated through mining these technical bibliographic databases via TOA,

TOA is an approach to mining information concerning the process of technological innovation. Based on a synthesis of the many models of technological innovation, diffusion, and transfer, we have identified candidate indicators of innovation status and prospects [3]. We seek empirical measures of those indicators through ‘Knowledge Discovery in Databases” (KDD) processes applied via an integrated suite of UNIX software tools known collectively as “TOAK” (TOA Knowbot). TOAK combines statistical, natural language processing, and fuzzy techniques with domain knowledge to perform quantitative content analyses of information contained in textual documents. The approach reflects an underlying belief that significant trends in technological research can be identified by sophisticated analysis of R&D documents databases. Such trends can then be used in assessing a technology’s potential for growth and related opportunities.

TOA is an instance of bibliometric analysis . In order to enhance effective science and technology management, various methods have been developed to understand the structural relationships among scientific papers or patent documents. These can yield “science indicators” – e.g.,, depicting R&D activity concentration and trends. Some bibliometric methods use quantitative techniques to extract and analyze information contained in text sources. One form examines citation patterns among papers or patents to detect seminal contributions, to ascertain interaction patterns across fields or institutions, and to forecast emerging research areas[19]. Co-citaflon analysis focuses on articles cited together in other articles [21] to discern cognitive linkages [20]. Co-word analysis, an alternative to citation and co-citation analysis, is also used to study how

* Ongoing TOA development is supported by the Defense Advanced Research Projects Agency (DARPA) and the U.S. Army Tank-Automotive Command (Contract DAAH01-98-X-R128). TOA is currently used to support analyses for industry and government, and a Windows version, TOAS, developed by Search Technology, Inc., Norcross, GA 30071-4701,USA,is in beta test [contact Paul-Frey for information- e-mail: paulf@searchtech.com]

documents or terms are related. Co-word analysis looks for words which co-occur disproportionately in documents. Co-word analysis can use either pre-identified index terms [l 8], or terms which emerge after techniques such as "stop word" removal and stemming have been used on a collection of documents. Kostoff has extended co-word analyses to whole text co-occurrence analyses [22,23,24]. These various bibliometric analyses often lead to sets of maps to portray the relationships among different science and technology domains (c.f., Kostoff's compilation at http://www.dtic.mil/dtic/kostoff/index.html).

Figure 2 tries to place TOA in perspective, In general, it draws upon broader information science and statistical techniques to elicit bibliometric information, Step-by-step, the TOA process first entails search in one or more databases, lending to retrieval of electronic document sets on the chosen topic. Then, basic data and text processing are applied to fuse records (if multiple data sets), remove duplicates, parse sentences into phrases, and to apply thesaurus and fuzzy matching techniques to consolidate related data. Inductive statistical analyses are then performed to group terms or documents using techniques such as Principal Components Analysis (PCA). Then information is represented graphically using various linking and clustering approaches such as multi-dimensional scaling. Finally, composite empirical indicators based on innovation models are provided.

"Knowledge discovery in databases" and "data mining"' are rapidly emerging fields. As such, they provide an excellent target to which to apply the TOA tool set.

FIGURE 1
Perspective on Bibliometrics

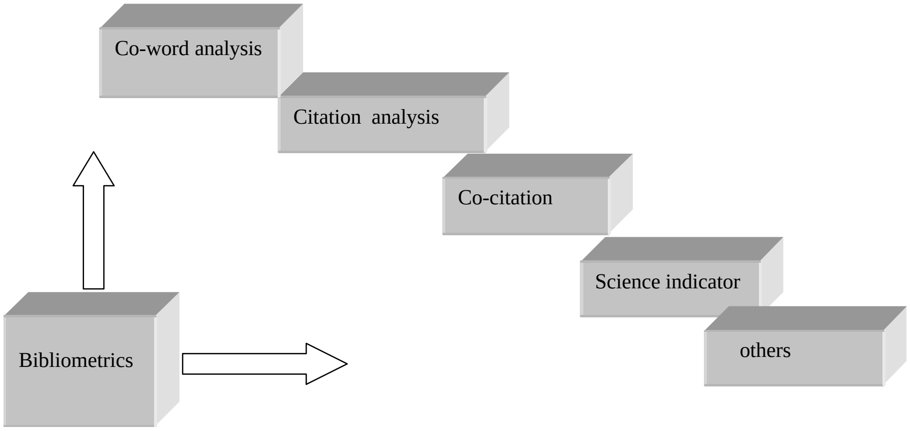
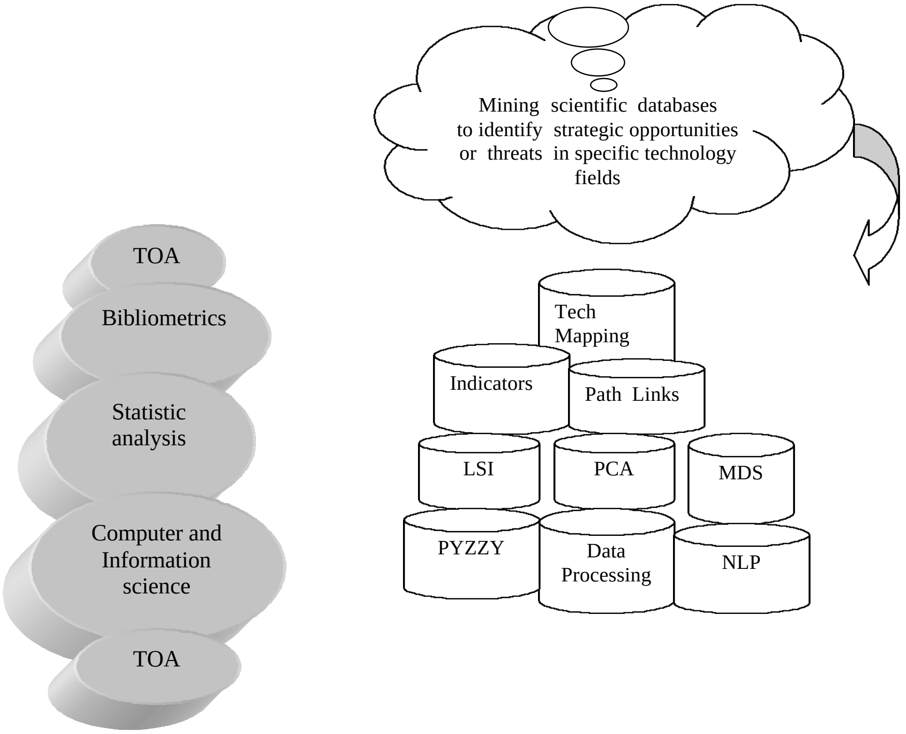


FIGURE 2
Perspective on Toa


Mining scientific databases to identify strategic opportunities or threats in specific technology fields
TOA
Bibliometrics
Statistic analysis
Computer and Information science
TOA
Tech Mapping
Indicators
Path Links
LSI
PCA
MDS
PYZZY
Data Processing
NLP

APPLICATION OF TOA TO THE CASE OF KNOWLEDGE DISCOVERY AND DATA MINING: BASICS

We have applied TOA to KDD and data mining, as a reflective effort to understand how TOA compares With others' approaches and to identify potential complementary tools. This paper illustrates the TOA process for a data set based on searches on "knowledge discovery" and "data mining."

"Data mining" ("DW") has evolved through long-standing, though varied, use by statisticians, data analysts, and ,management information systems specialists [5, 10]. We explore its prevalence in various databases (Table 1). We pursue analysis particularly in *INSPEC,* a rich, research-oriented source database, for abstracts containing "data adjacent to mining" from 1986-97. *(COMP, The Computing Index.* is more oriented to trade magazines.)

TABLE 1
Data Mining Documents in Databases (search term: data adj. mining)

Databases	INSPEC	ENGI	BUSI	COMP	NTIS	US PATS	GTEC
# of documents	575	225	142	680	78	2	11

ENG= Engineering Index; BUSI= Business Index; NTIS= National Technical Information Service; US PATS= US Patents; GTEC= Georgia Tech Library Catalog.

In exploring the DM data set, we noted the extensive use of "knowledge discovery." This led us to search on knowledge discovery in databases (KDD) in INSPEC, yielding 307 records. Of those, 188 appear in common With the DM data set.

Tijssen and Van Raan have distinguished one-dimensional (list-based) and two-dimensional (matrix-based) analyses in bibliometrics [25]. Narin [1994] categorizes bibliometric methods as l) activity measures (e.g., counts of publications or patents, by topical area or institution), 2) impact measures (e.g., citations), and 3) linkage measures (e.g., evidence of intellectual associations).[17] We begin with activity counts in this section, then shift toward 2dimensional linkage measures in the following sections.

Basic counts of keywords (subject index terms) in the KDD and DM data sets show that the top 50 keywords in KDD closely match those in DM abstracts. Furthermore, the top 20 authors and top institutions contributing to KDD are quite similar to those in DM. (In another analysis of the conceptually close fields of natural language processing and computational linguistics, we found striking dissimilarity between the respective author and institutional contributors [4]) We then considered what particular papers said about the relation between KDD and DM, deciding that the relationship is indeed close [5]. We therefore consolidated the KDD and DM abstract sets into a unified search set of 694 records upon which the remainder of this paper focuses.

There are many authors contributing to this KDD/DM literature. After applying TOA fuzzy and thesaurus routines, with our review, to combine variants of the same name, we identify 1142 authors for these 694 articles – obviously multiple authorship is common practice. The appendix shows the 23 most prolific authors, with 6 or more KDD/DM publications each.

We next pursue the institutional affiliations of the authors. In all, 398 affiliations are identified; those associated with the most publications are listed in the appendix. Some 27% of the contributions are from companies, led by IBM. The interests of specific companies could be probed further by searching for websites (e.g., http://www.almaden.ibm.com/almaden/).

One can pursue these explorations many ways. For instance, inspection of the affiliations of the most prolific authors shows a similar split between company associations (22%) and academic (74%). This suggests that KDD/DM research is predominantly academic, but with a significant industry involvement. On an individual level, one could explore the worldwide web to note that Fayyad has moved from Caltech's JPL to Microsoft. Such exploration offers another angle on the backgrounds relating to KDD/DM. For instance, Mannila [6] notes that KDD/DM draws upon machine learning, statistics, and databases; Klemettinen et al. [11] point toward the interface of computer science and statistics in KDD/DM. We have linked our homepage [http://tpac.gcatt.gatech.edu] to a number of the leading authors' homepages. Their self-descriptions suggest strong ties to machine learning (for 9 of the 23 prolific author), database systems (8), and various artificial intelligence areas (5).

These basic analyses help ascertain the emphases of a field — KDD/DM in this case — and who the active players are, They provide a nice introduction to "who's doing what" that can be extended through direct interactions with active researchers in the field, We now turn to "two-dimensionally" analyses that seek to uncover relationship mining the database information further.

DOMAIN MAPPING THROUGH CLUSTERING TECHNIQUES

TOA performs co-word analysis to infer underlying relationships in text documents. Such analyses are very useful to the analysis of bibliographic data in the form of field-structured abstract records, in particular, and often focus on keywords. In TOA, we seek first to cluster empirically related terms and then to depict relationships as "technology maps."

The TOA software provides capabilities to 'mix and match' a number of grouping and linking algorithms. In addition we have experimented with a wide range of these, generally speaking, 'clustering' approaches. One can examine co-occurrences among various types of terms – for instance, keywords by authors, or affiliations by year, or whatever is of interest. Here, we focus on keywords by keywords, seeking insights into how these terms conceptually cluster together. One can group these co-occurring entities by such techniques as Latent Semantic Indexing [26, 27], Principal Components Analysis (PCA), factor analysis, hierarchical cluster analysis, or Maximum Likelihood Intensity Similarity analysis [28]. Then, one may want to graphically represent relationships. Techniques such as spanning trees, Pathfinder [29], multi-dimensional scaling, or path-erasing [12] come into play.

We illustrate the genre with a particular two-stage Term Cluster Mapping approach [2]. Basically, the approach first clusters keywords based on the co-occurrence patterns, then maps these showing links among those clusters- The term cluster mapping process entails three steps (Figure 3).

(1) identify the term clusters;
(2) build a similarity matrix of the clusters;
(3) represent the similarity matrix in low-dimensional mapping.

For Step 1, we chose 106 keywords most prominent (occurring 5 or more times) in the KDD/DM data set excluding the search terms and direct derivatives of them. We applied PCA — a well-recognized statistical procedure that generals linear combinations of the input variables (in this case, the occurrence pattern of the 106 keywords across the 694 documents), such that the first such “factor” explains the most possible variance; the second factor, the most possible remaining variance; and so forth. We extracted 30 such factors, than rotated the factors so that keywords would tend to relate either highly or not to each factor (Varimax rotation). We then applied a heuristic to identify the keywords relating closely to each factor — so-called “high loading” terms. These are listed in the appendix.

FIGURE 3
The term clusters mapping process

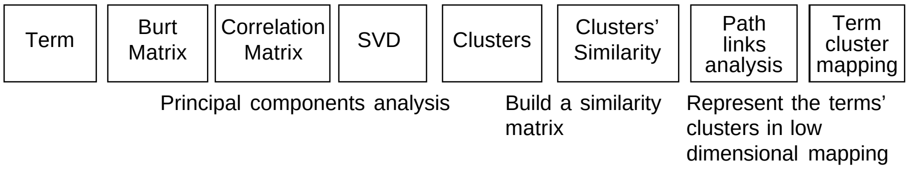
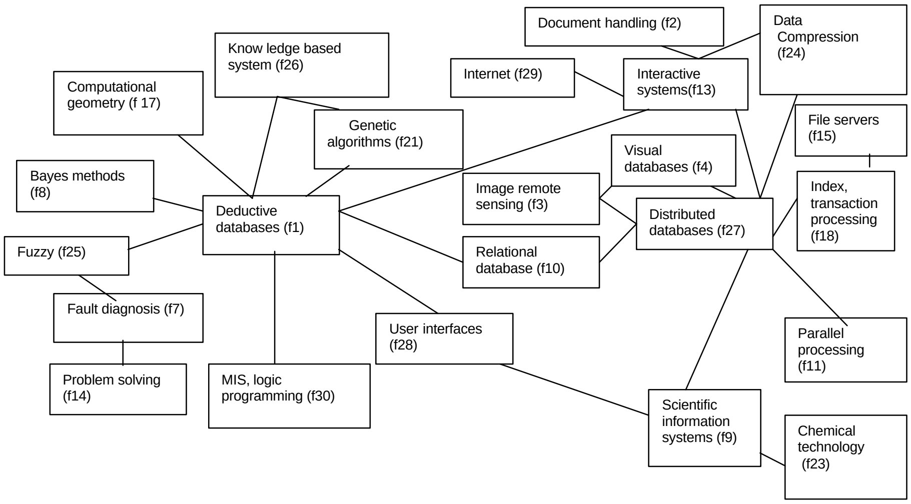

Term
Burt Matrix
Correlation Matrix
SVD
Clusters
Clusters’ Similarity
Path links analysis
Term cluster mapping
Principal components analysis
Build a similarity matrix
Represent the terms’ clusters in low dimensional mapping

FIGURE 4
Term’s cluster mapping


Document handling (f2)
Data Compression (f24)
Know ledge based system (f26)
Internet (f29)
Interactive systems(f13)
Computational geometry (f 17)
Genetic algorithms (f21)
File servers (f15)
Visual databases (f4)
Bayes methods (f8)
Image remote sensing (f3)
Index, transaction processing (f18)
Deductive databases (f1)
Distributed databases (f27)
Fuzzy (f25)
Relational database (f10)
Fault diagnosis (f7)
User interfaces (f28)
Parallel processing (f11)
Problem solving (f14)
MIS, logic programming (f30)
Scientific information systems (f9)
Chemical technology (f23)

For Step 2, one must decide how to gauge the similarity o 1 f the clusters (the factors). We used a group-average method. For example, consider Factors 4 and 5. We average the Pearson correlation between “visual databases” (one of the two high-loading terms of Factor 4) and “decision theory” (¡.e., a normalized co-occurrence measure of how often the two terms appear together in documents); “visual databases” and “trees”; ‘spatial data structures” (the other high-loading term of Factor 4) and “decision theory”; and ,”spatial data structures” and “trees” (“decision theory” and “trees” being the high-loading terms-of Factor 5). Carried out for all 30 factors, this yields the similarity matrix of clusters (factors).

Step 3 consists of representing these similarities among factors in a two-dimensional map (Figure 4). Positioning in the map is relative, representing which factors are empirically most central to this data set (KDD/DM), where centrality is based primary upon the strength of similarities to the set of factors. Links shown are based on our “Path Erasing” approach. This is an algorithm that begins with each entity linked to every other, then removes links to a designated threshold level. The level can be adjusted; the intent is to convey the main relationships. (In other words, the absence of a link does not indicate total independence, rather notabiy less association.) Figure 4 shows *‘Deductive Databases” and ‘Distributed Databases”* as prominent centers, themselves linking particularly through *‘Relational Databases,”* Note that these are factors, not individual terms. *“Deductive Database”* our name for Factor 1, composed of two terms – “knowledge acquisition’ and “deductive databases.” In other words, Figure 4 is showing relations among clusters of keywords, not among the keywords per se.

Examination of the similarity matrix of factors, leads us to suggest two major domains within this KDD/DM research. Domain 'A' centers, around *'Deductive Databases '* Some of the individual keywords (Note these are not shown in Figure 4 which only shows keyword clusters) correlating strongly include: very large databases, inference mechanisms, generalization, database theory, machine learning, inheritance, pattern recognition, interactive systems, relational database, logic programming, uncertainty handling, information theory, genetic algorithms, etc.,

(called group A). These *seems* to us to be quite research oriented. The other domain, "B," associates with *'Distributed Databases.'* This includes such individual keywords (not shown in Figure 4) as: object-oriented, geophysics computing, information networks, data structures, multimedia, query, transaction processing, parallel processing, decision support systems, file servers, visual databases, internet, data acquisition, digital simulation, iterative methods, data reduction. These seem more application-oriented in our opinion.

The high-loading terms of *"Deductive Databases" are* much more prevalent than those of *"Distributed Databases"*

(238 vs. 34 documents represented, of 694),.) it suggests that "Deductive Databases' is a hot area in KDD and data mining, We also found that very large databases is the closest term to deductive databases, it suggests that deductive data bases should have strong and important applications in the future of KDD and data mining although it is still research driven right now. Domain B – *"Distributed Databases"* – appears to be an

emerging emphasis for KDD/DM, possibly worth special attention by the field.

TECHNOLOGY PROFILING USING MULTIPLE INDICATORS

We introduce a set of derived indicators intended to help understand the status and prospects for ongoing innovation (development toward application) of the topic under study – in this case. KDD/DM. These include:

* A.n.a.p – the average number of authors per paper, measuring the size of research teams;

* Rc.j – the ratio of conference papers to journal papers, indicating rapidity of communication.

A set of indicators has been developed and introduced in the TOA process. For example,

* Tn.p, the total number of papers of a domain, describes the size of a domain;

* Ar.p.p is the growth rate, which is the average growth of a domain's documents in a certain period;

* Pc.p.p, the percentage of companies' papers in all papers of a domain, portraying the extent of industrial interest:

* Na.m, which is a domain or term's normalized association measure with a given data set in a certain period, measuring the domain/term's association with a given data set.

The KDD/DM data set averages 3,23 authors per paper. A high A.n.a.p in a data set implies large research teams and that suggests application, or experimental, or interdisciplinary group research. A low A.n.p.p suggests more individualistic, theoretical efforts. In order to make a comparison, we retrieved three other data sets from the same INSPEC database source: 'Database Systems,' 'Machine Learning,' and 'Statistical Analysis.' These, respectively, average 3.43, 3.18, and 2.94 authors per paper. KDD/DM collaboration patterns lie in the middle range, quite similar to those for 'Machine Learning.'

KDD/DM researchers are more inclined to present their work via conferences than through archival journals. The Rc.j for the KDD/DM data set is notably high – 3.0. This compares with 0.506 for "Statistical Analysis'; 0.578 for 'Decision Theory'; 5.28 for 'Spatial Databases'; 4.64 for 'Association Rules'; 2.46 for 'Visual Databases'; 2.39 for "Deductive Databases'; and 2.35 for 'Genetic Algorithms.' We interpret this to suggest that KDD/DM is a relatively 'hot' research area in which speed is of the essence in dissemination of results. We have devised a number of such indicators – others are introduced below in composites; also see an extended version of this KDD/DM analysis on our website (http://tpac.gcatt.gatech.edu). We move on to explore 'two-dimensional' indicators.

Size vs. Growth Rate

INSPEC keywords provide one way to get at related technical topics. We focus now on the leading 36 keywords which are higher frequency terms in the KDD/DM data set. (One might prefer to use the 30 factors, composites of the keywords, but we chose the pre-established individual keywords to keep this analysis dearer.) Two key aspects of technical topics in characterizing a research domain are the Size (total activity) and Growth Rate (change in that activity over time). For each of these 36 keywords, we examine their profile in the overall INSPEC database (containing wm 3,000,000 abstracts for the period 1987-97, concentrating on R&D in computing, electrical engineering, and the physical sciences). We group them into eight categories based on Intensity and Growth Rate, the most interesting being:

• Small Size; extremely high Growth Rate — 1 term: worldwide web

• Small Size; high Growth Rate – 10 terms; (1) data visualization; (2) association rules; (3) business data processing; (4) visual databases; (5) rough sets; (6) very large databases (7) data warehouse; (8) spatial databases; (9) pattern classification; (10)query processing

• Medium Size; high Growth Rate – 1 term: genetic algorithms

Focusing on these 12 terms that show particularly compelling growth, we now break out according to the extent of industry involvement in R&D for each. The rationale is that increasing industry involvement reflects increasing commercialization opportunity. Figure 5 locates each technical area by its growth rate (Y axis) and industrial involvement (X axis). We observe:

I) "Data warehouse" is apparently mainly driven by companies at present.

II) Research on "Association rules,' 'business data processing" and "very large databases' (group II in Figure 5) is driven by both companies and academia units, but companies are notably active in these three areas.

III) "Data visualization," "visual databases,' "spatial databases," and 'KDD/data mining' are pushed by both industrial and academic units,

IV) "Rough sets," "genetic algorithms," and 'pattern classification" are pushed largely by academic units right now.

V) 'World wide web' shows an extremely high growth rate, with both companies and academic units paying it much attention.

Examination of the full set of 36 technical areas prominently associated with KDD/DM articles yields additional insights. First, there is a general trend for those topics engaging industry researchers to be fast growing (not shown here; see extended paper on the website). Two groups of technical areas deviate from this general trend. The retrieval, relational databases, expert systems, and knowledge based systems. This suggests possible maturing of the field or slowdown in terms of research, but possible "maturing" toward commercial potential (determination of which would require additional review), The second deviant group (IV, from Figure 5) shows relatively high growth (Within this group "Rough sets" is still small so might well attract industry attention if activity increases.)

FIGURE 5
The development patterns of the domains with high growth rate

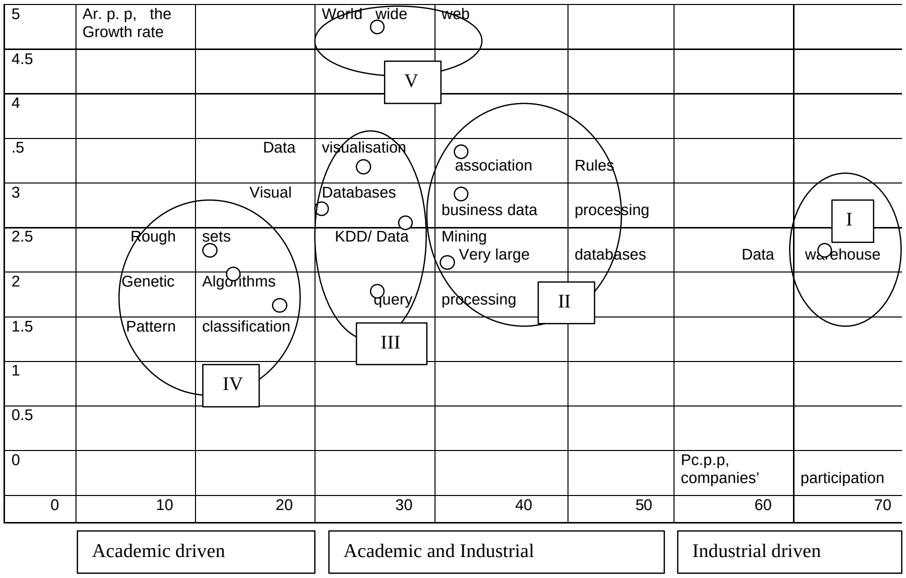
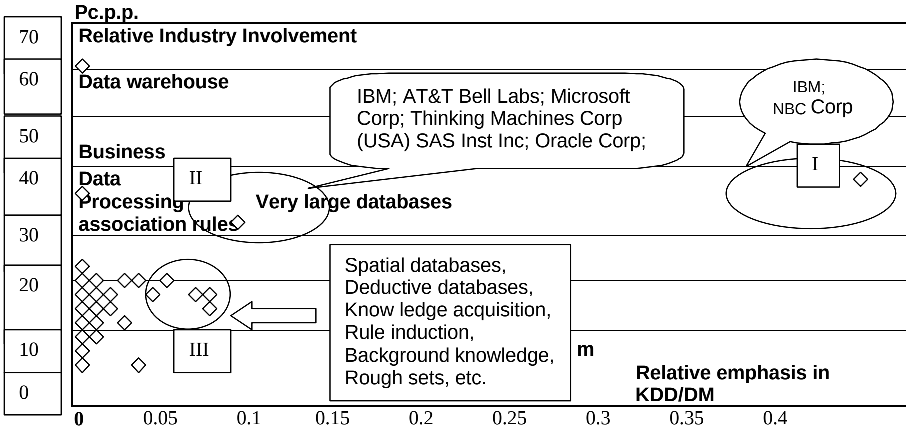


FIGURE 6
What are the hot areas with heavy company involvement



Topical Specialization

Relative emphasis is the ratio of a keyword's occurrences in KDD/DM documents to the keywords occurrences in the source database (*INSPEC). In* other words, a higher ratio implies that the technical area is relatively particular to KDD/DM. The relative emphasis indicator shows that association rules, very large databases, deductive databases, knowledge acquisition, rule induction, spatial database, background knowledge, rough sets, etc., are relatively particular to KDD/DM at present. Except for the term 'knowledge acquisition," the top ten hot fields in KDD and data mining are small in size. This fact suggests that there is still a very broad space for fresh corners to take part in KDD and data mining. It seems that people may not have to worry who is ahead of us if they want to join the competition of KDD and data mining since the race is just beginning,

Figure 6 locates the 36 technical areas, as just discussed, in terms of relative industry involvement (Y axis) and relative emphasis in KDD/DM (X axis).

The two technical areas most concentrated in the KDD/DM domain, "association rules" and "very large databases" both show specially strong industry participation. IBM is notably active in publishing on 'association rules'; a number of companies are publishing aggressively on 'very large databases' (IBM, AT&T, Microsoft, Thinking Machines, SAS Institute Oracle, MCC). It's interesting that the remaining KDD/DM technical areas are mainly academic, with the exception of "business data processing' and the striking outlier, 'data warehouse.' This suggests that many of the KDD/DM basic approaches/techniques are still predominantly being addressed in academia. Industry might want to track developments in these domains with special attention to identify early opportunities for commercial application.

INTERPRETATION

This paper reflects application of a combined bibliometric/language processing approach, TOA, to profile research in KDD/DM. We can summarize the TOA, to process briefly as follows:

- Information Retrieval – Boolean search in bibliographic databases; abstracts passed to the TOA software for cleaning (duplicate removal, fuzzy matching, etc.) and analysis;

- Knowledge Acquisition – Primarily statistical approach operating in what could be cast as a vector space model of terms by documents

- Information Organization & Representation – Data reduction, mainly linear algebraic;

- Adaptive partitional clustering (e.g. ,bucketing of concepts or documents – not illustrated here), and network concept linking.

The application of TOA to scientific and technical document databases is illustrated for KDD and data mining. Outputs of TOA include basic profiling of the R&D activity in the target area, mapping of technical topic interrelationships, and composite "science or technology indicators"[33]. A modest subset of potential indicators has been presented. Many of these can be extended interestingly over time (e.g., to yield time slices indicating changes in topical emphasis or industry involvement?). Others can be pursued to another level of detail (e.g., repeating these analyses on subsets addressing only the work of a particular company, or on a particular sub-topic. to yield competitive intelligence; The indicators presented emphasize the main activities — the "usual.". Another set of indicators can spotlight the 'unusual' (e.g., instead of focusing on the prominent keywords or authors, focus on the novel abstract phrases or authors making their first contribution to the target domain). We have also experimented with reproducing TOA's on topics, say, 6 months later to highlight 'what's new.' We believe such analyses hold potential value for a wide range of professionals with a stake in science and technology, for instance:

- Researchers (e.g., to identify relevant work outside their normal domain of inquiry and network)

- R&D managers (e.g., to familiarize themselves with alternative approaches being applied in a domain of interest

- Technology information specialists, concerned with technology licensing, competitive technical intelligence, etc. (e.g., to identify threats and opportunities in a technology)

- Strategic planners and managers (e.g., to benchmark leaders in an area and determine whether to pursue that area).

TOA and similar analyses can help newcomers, such as students, understand what is involved in the development of a technical field, such as KDD. They can help experienced players in the field gain new perspectives that can uncover gaps, suggest new opportunities to apply their strengths, and pose possible linkages beyond their normal span of interests. TOA, or other bibliometrics, do not provide the 'final word,' rather, they offer a different viewpoint that can instigate more detailed investigations.

REFERENCES

1. Porter, A. L., & Detampel, M.J., Technology Opportunities Analysis**,** *Technological Forecasting and Social Change* No.49, 237-255, 1995.

2. Porter, A. L., etc., (1998). Georgia Tech Research & Development TOA, *Final Report on subcontract to Search Technology, Inc.,* for Defense Advanced Research Projects Agency STTR Phase 2 Project, Technology Opportunities Analysis System [DAAHO 1 -96-C-R 1 691.

3. Watts, R.J., and Porter, A.L., innovation Forecasting, *Technological Forecasting and Social Change* No.56, 2@47 1997.

4. Watts, R.J., and Porter, A.L., Cunningham, S.W., and Zhu, D., TOAS intelligence Mining; Analysis of Natural Processing & Computational Linguistics, in Komorowski, J. and Zytkow, J.(Eds.), *Principles of Data Mining and Knowledge Discovery,* Spinger, 1997.

5. Usama Frayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, knowledge Discovery and Data Mining: Towards a Unifying Framework, *Proceeding of the Second International* Conference on *Knowledge* Discovery *and Data Mining* (KDD-96), Portland, Oregon, August, 1996., AAAI Press.

6. Heikki Mannila, Data mining: machine learning, statistics, and databases, *Eight International Conference on Scientific and Statistical* Database *Management,* Stockholm June 1996, p. 1 -**8.**

7. Karsten M. Decker, Sergio Focardi, Technology Overview: A Report on Data Mining, *Technical Report,* (CSCS TR-95-02) Swiss Scientific Computing Center, 1995.

8 Ming-Syan. Chen, Jiawei Han, and Phiilip S. Yu, Data Mining: An Overview from Database Perspective’, IEEE *Transactions on Knowledge and Data Engineering, 1997.*

9. Marcel Holsheimer, Marün Kersten, Heikki Mannila, Hannu Toivonen, A perspective on Databases and Data Mining, *The First International Conference on Knowledge Discovery and Data Mining,* Montreal, 1995.

10. Clark Giymour, David Madigan, Daryl Pregibon, Padhraic Smyth, Statistical Themes and Lessons for Data Mining, *Data Mining and Knowledge Discovery, 1,* 25-42, 1996.

11. Mika Kiemeftinen, Heikki Mannila, Pirjo Ronkainen, Hannu Toivonen, A. lnkeri Verkamo, Finding Interesting Rules from Large Sets of Discovered Association Rules, *The Third International Confer Information and Knowledge Management,* Ed. Nabil R. Adam, Bharat K. Bhargava and Yelena Yesha, 404-407. Nov., 1994, Gaithersburg, Maryland. ACM Press.

12. Stefan Wrobei, Dietrich Wettschereck. lnkeri Verkamo. Arno Siebes, et., User Interactive in Very Large Scale Data Mining, Pro.-. *FGML -96. Annual Workshop) of the* L’71 *Special Interest Group Machine Learning* (GI FG 1.13), lnforrnatik-Berichte, Univ. Chemnitz-Zwickau, 1996.

13. Hekki Mannila, Hannu Toivonen, A.Inkeri Verkamo, improved Methods for Finding Association Rules, *Tech* Reporter, University of Heisinki, Department of Computer Science, Series of Publications C, No. C-1 993-65

14. Marcel Holsheimer, Arno Siebes, Data Mining: the Search for Knowledge in Databases, *Report CS-R94O6Í ISSN 0169-ll8X,* CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands.

15. Stefan Wrobel, Saso Dzeroski, The ILP Description Learning Problem: Towards a General Model-Level Definition of Data Mining in ILP, *Proc, FGML-95, Annual Workshop of the GI Special Interest Group Machine Learning* (Gl FG 1. 1.3), ed. K. Morik and J. Herrmann, Research Report 580, Univ. Dortmund, 1995.

16. Ezawa, K.J.; Norton, S.W,, Knowledge discovery in telecommunication services data using Bayesian network, in Fayyad, U.M. and Uthurusamy, R.(Eds.), *KDD-95 Proceedings. First International Conference on Knowledge Discovery and Data Mining, 1995.*

17. Narin, F., Olivastro, D., and Stevens, K.A., bibliometrics-Theory, Practice and Problem, *Evaluation Review* 18(1), 1994.

18. Calion, M., Courüai, J. P., Crance, P., Laredo, P., Mauguin, P., Rabehaisoa, V., Rocher, Y. A., and Vinck, D., Tools for the Evaluation of Technological Programmers: An Account of Work Done at the Centers for the Sociology of 1 innovation, *Technology Analysis and Strategic Management* 3(1), 3-41, 1991.

19. Garfield, E., Malin, M.V., and Smali, H., Citation Data as Science indicators, *The Metric of Science: The Advent of Science Indicators,* Y. Elkana et al., eds., Wiley, New York, 1978,

20. Melkérs, J., bibliometrics as a Tool for Analysis of R&D impact, in Evaluating R&D impacts: Methods and Pracüce, B. Bozeman and J. Melkers, eds., Kluwer, Boston, 1993, pp. 43-61.

21. Smali, H., and Griffith, b., The Structure of Scientific Literatures, *Science Studies* 4, 17-40, (1974).

22. Kostoff, R.N., Research impact Assessment: Problems, Progress, Promise, *Proceedings: Fourth International Conference on Management of Technology,* Miami FL, 1994.

23. Kostoff, R. N., @-Word Analysis, in Assessing R&D impacts: Method and Practice, Bozeman, B. and Melkers, J., Eds. (Kluwer Academic Publishers, Norweil, MA) 1993.

24. Kostoff, R. N., Database Tomography for Technical intelligence, *Competitive Intelligence Review, 4:* 1, Spring 1993.

25. Tiissen-RJW and Van Raan-AFJ, Mapping Changes in Science and technology - Bibliometric Cooccurrence Analysis of the R-and-D Literature, *EVALUATION REVIEW,* Vol 18, lss 1, pp 98-115.1994.

26. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, D. (1990). indexing **by** Latent Semantic Analysis, *Journal of the American Society for Information Science 41 (6),* 391-407.

27. Dumais, S. T., Furnas, G. W., Landauer, T. K., & Deerwester, S. (1 988). Using latent semantic analysis to improve information retrieval. *Proceedings of CHI’88: Conference on Human Factors in Computing.* New York: ACNÍ,, 281-285.

28. Cunningharr, S. (in submission), A Maximum likelihood Approach to the Mapping of Text, *Journal of the American Society for information Science.*

29. Roger W, Schvaneveldt (Editor), Pathfinder Associative Networks: Studies in Knowledge Organization, Published 1990. ISBN 0-89391-624-2

30. National Science Board(1998). Science *& Engineering Indicators* - 1998. Arlington, VA, USA: National Science Foundation.

31. Ministerio da Ciencia e Tecnologia(1996). *Indícadores Nacionaís de Ciencia & Tecnología, 1990-95.* Brasilia:MCT.

32. Roessner, J. D., Porter, A. L., and Newman, N. C. (1 997). *1996 Indicators of Technology-based Competitiveness of Nations.* Atlanta GA, USA: Technology Policy and Assessment Center, Georgia Tech (report to the National Science Foundation under D22588X).

33. Newman, N.C., Porter, A.L., and Cunningham, S.W. (1 997). ‘technology Opportunities Analysis for Malaysia,’ *Portland International Conference on Management of Engineering and Technology,* Portland, OR, USA {CD}.

Donghua Zhu

Zhudh@isye.gatech.edu

Alan Porter

alan.porter@isye.gatech.edu

Scott Cunningham
Judith Carlisie
Anustup Nayak

Technology Policy & Assessment Center,
Georgia Institute of Technology, Atlanta, GA
30332-0205