Data Min Knowl Disc (2007) 15:99–105
DOI 10.1007/s10618-006-0058-2

Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from "university" to "business" and "analytics"

Gregory Piatetsky-Shapiro

Received: 12 May 2006 / Accepted: 25 September 2006 / Published online: 27 January 2007

Abstract I survey the transformation of the data mining and knowledge discovery field over the last 10 years from the unique vantage point of KDnuggets as a leading chronicler of the field. Analysis of the most frequent words in KDnuggets News leads to revealing observations.

Keywords Hype curve · Business analytics · KDnuggets · Data mining jobs

1 The big picture

Over the last 10 years the data mining and knowledge discovery field went through enormous transformation, influenced by seismic external forces such as the enormous growth of web/e-commerce, tremendous progress in biology (such as DNA microarrays), and frequently controversial use of data mining for homeland security. New research areas have emerged such as social network (link) analysis, web mining, and multi-media mining. Tremendous progress was made in refining association rules algorithms (although good applications are still few), and new important algorithms such as SVM have become widely accepted. Open-source data mining software such as Weka and R led to wider use of data mining tools.

In 1996, there was one conference (KDD-96) and about 100 research papers dealing with data mining. In 2005, there were over 20 conferences (see www.kdnuggets.com/meetings/past-meetings-2005.html), with several thousand

Responsible editor: Geoffrey Webb

G. Piatetsky-Shapiro
KDnuggets, 22 Atherton Road, Brookline, MA 2143, United States
e-mail: gregory@kdnuggets.com

papers. Data mining research and industry became more international, with strong growth in Asia (especially in China).

However, there are too many changes to summarize in a short survey paper and therefore I will focus on my unique perspective as the editor, since 1993, of KDnuggets News (www.kdnuggets.com/news/), which was and still is the leading chronicler of the data mining community.

For another retrospective from 1989 to 1999 see (KDD-99 Panel Report).

(Note: In this paper, I will follow common usage and be using terms "Data Mining" and "Knowledge Discovery" as synonyms.)

2 Data mining: overcoming the Hype

New technology progress frequently follows the aptly named "Hype Curve" which compares the technology's actual performance versus expectations (Eric Brethenoux, Personal communication). When a new technology first appears, the expectations are modest, but if it shows promise, then expectations begin to increase — typically much faster than the actual performance, leading to a "hype peak" of over-inflated expectations. The big gap between high expectations and actual performance then leads to a "trough of disappointment" where expectations actually fall below actual performance (we saw that happen in early 1990s during the "AI Winter"). However, if technology has real promise, it recovers, reaches growing acceptance, and becomes mainstream.

Figure 1 shows the Hype Curve for Knowledge Discovery I first presented, surprisingly, back in 1996.

This curve described the progress of our field quite well, although, like most people, I was wrong on the timing of the "hype peak". The peak of expectations was not in 1997, as shown on the graph, but in the fall of 1999 — spring 2000, when there were a number of data mining related IPOs and acquisitions for what seemed then like an astronomical amount of money, including

- E.piphany bought Rightpoint for around $ 400 million, November 1999 (KDnuggets News 99: n24).
- Vignette bought DataSage for $ 555 million, January 2000 (KDnuggets News 2000:n01).

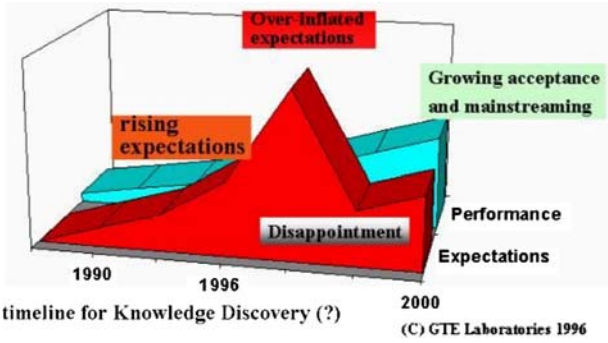

Fig. 1 Hype curve for knowledge discovery, presented in 1996

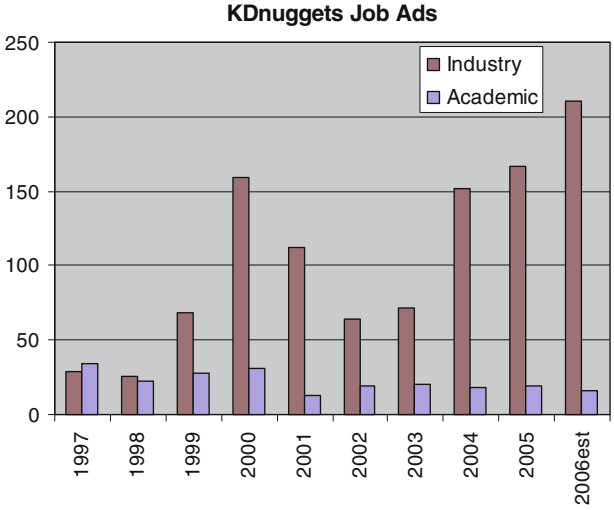

Fig. 2 KDnuggets job ads: industry versus academic/government

- Xchange bought Knowledge Stream Partners for $ 52 million, April 2000 (KDnuggets News 2000:n07).

This hype peak was then quickly followed by dot com collapse in summer and fall of 2000. Many dot com and data mining-related companies (including those mentioned above) lost almost all of their value, and some companies went out of business. However, around 2002, the bottom was reached and recovery began in 2003.

One approximate but useful measure of demand for data mining is the number of job/position ads in KDnuggets (Fig. 2), broken down into Industry and Academic/Government ads. We see that back in 1997 there were more academic than the industry jobs, but by 2005 the academic positions have declined somewhat while the industry jobs have skyrocketed. The industry job growth also corresponds well with the hype curve above — we see the "hype" peak in 2000, followed by decline and recovery starting in 2003, and surpassing the previous high in 2005. The number of job ads for 2006 was estimated by taking the data for the first two quarters and annualizing it.

Note the record increase in industry job ads in 2006 (especially noticeable in quarterly data not shown here). Based on our experience, it is likely to indicate another "hype peak," after which a temporary decline to more sustainable job demand rate is likely to follow in 2007.

3 Trend analysis: top Keywords in 1996 and 2005

I have been publishing KDnuggets News (www.kdnuggets.com/news/) since 1993. Based on the number of subscribers and search engine ranking for www.kdnuggets.com, KDnuggets News can be considered a leading forum for

Table 1 Top 10 words in KDnuggets News in 1996 and 2005

Rank	Words 1996	Count	Words 2005	Count
1	Data	2571	Data	3545
2	Mining	1385	Mining	2166
3	University	898	Com	1822
4	Com	860	Business	756
5	Information	856	Experience	702
6	Knowledge	833	Research	685
7	Research	723	Analysis	669
8	Kdd	674	Software	665
9	Edu	636	Information	629
10	Discovery	623	Modeling	498

data mining community. It is therefore illuminating to examine the changes in terms and keywords used in KDnuggets News. While this is a simple text mining exercise, the results are very revealing. (Note: more complex analysis of word pairs and frequent itemsets is possible, but I did not have time to do it. It is a good project for interested text miners and students - all KDnuggets News issues are available on the web).

I extracted all words from KDnuggets News in 1996 and in 2005. After removing HTML tags, punctuation, and common words, but keeping words in URL and email addresses, the top 10 most frequent words in 1996 and in 2005 year appear in Table 1.

The two most common words are, unsurprisingly, "data" and "mining." However, the third most common word in 1996 was "university", reflecting the university base of data mining activity then. The most significant new word in 2005 is "business," reflecting the business orientation of the data mining field. We also see more than doubling of string "com" which almost always corresponds to ".com" domain (from 860 to 1,822) and decline of ".edu" domain frequency, confirming the shift from university to business applications.

The word "experience" is most commonly used in job ads. We see that "research" is still very much present in 2005 KDnuggets News, but "software" became more important.

We next compared the words that became more prominent in 2005 compared to 1996 (limited to words that appeared at least 50 times in 2005).

We note that "analytics" increased in frequency by 24,600% ! This, together with the prominence of the "business" word represents the shift of discussion from university research to business analytics (Table 2).

Additional prominent words represent a new feature of KDnuggets (poll), SIGKDD — the ACM data mining and knowledge discovery organization, several new companies prominently involved in data mining (Yahoo, Insightful, Amazon, Providian, KXEN, Google, Equbits), existing companies that became much more prominent (SPSS), and a popular location for data mining conferences (Las Vegas) (Table 3).

Table 2 Top 10 words that most increased in frequency in 2005

word2005	Count 96	Count 2005	% Increase
Analytics	2	492	24,600
Yahoo	1	177	17,700
Insightful	1	146	14,600
Analytic	2	148	7,400
Agency	3	177	5,900
Statistician	1	52	5,200
Crm	2	78	3,900
Innovation	2	77	3,850
SPSS	4	151	3,775
Bi	2	61	3,050

Table 3 Top 10 new words in 2005 that did not appear in 1996

Word	Count 2005
Poll	180
Sigkdd	176
Amazon	159
Providian	129
Kxen	105
Google	103
Vegas	100
Kdd2005	99
Equbits	63
Engine	62

Table 4 Words from 1996 KDnuggets news with the largest drop in 2005

Word	Count 1996	Count 2005	Decline	Comment
Ftp	134	2	98.5%	Decline in FTP after web arrival
Almaden	63	1	98.4%	Decline of IBM Almaden as a major center for research in data mining
Postscript	61	1	98.4%	Decline of postscript as a publication format
Informatik	59	1	98.3%	Informatik was used at several german universities as part of web address
Rough	109	2	98.2%	Decline in popularity of rough set research

4 Comparing 2006 and 2005

Data miners are always trying to predict the future (Table 4). One possible indicator of current trends is comparing the keywords in 2006 to those of 2005.

I computed keyword frequencies for 2006, and annualized them (since only 17 issues were published in 2006 so far compared to 24 issues in 2005). I further required keywords to have annualized 2006 count $\geq$24 and appear in three or more items. Here are the keywords with the most significant increase, as measured by a simple chi-square test (Tables 5, 6, 7).

Table 5 Words from 1996 KDnuggets news not present in 2005

Word	Count	Comment
GTE	363	I have left GTE labs and GTE is now part of Verizon
Moderator	84	Moderator term is no longer used
Siftware	80	This cute term for Data Mining software no longer used
GMD	78	GMD no longer has an active group in data mining
JAIR	67	JAIR is not publishing articles on data mining

Table 6 Words with largest increase in 2006 KDnuggets news compared to 2005 (ordered by significance)

Word	Count 06 annualized	Count 2005	Comment
nationwide	36.7	1	company which started to hire data miners
ODM	52.2	3	Oracle Data Mining tutorial
NSA	73.4	6	NSA data mining in the news
burbank	28.2	1	Yahoo hiring in Burbank, CA
Unica	24	1	Unica buys MarketSoft, Sane
Monash	24	1	fellowships; Geoff Webb at Monash
AOL	81.9	11	AOL search data release
Vioxx	45.2	4	Analysis of Vioxx risks - was there a mistake?
KDD2006	57.9	7	KDD 2006
Prudsys	35.3	3	data mining company in the news
Effects	28.2	2	various stories
Philadelphia	66.4	10	KDD 2006 in Philadelphia
HP	32.5	3	HP Labs hiring
Merck	24	2	Merck Vioxx data analysis
Advertising	86.1	19	multiple stories related to advertising
Treenet	55.1	9	Salford Systems software
Blog	50.8	8	blogs related to data mining
Regulatory	25.4	3	stories both on financial regulatory issues and genetic regulatory networks
Inrix	25.4	3	Company which started to hire data miners
Microsoft	259.8	123	many stories, with microsoft becoming much more active in data mining

5 Summary

To summarize, over the last 10 years the data mining field went through a great transition. The field has recovered from the hype peak of the dot com era and is now becoming part of the mainstream business process. Business analytics (including internet businesses such as Amazon, Google, and Yahoo) is presenting the greatest demand on data mining skills and methods.

Having been through one hype curve we need to be wary of the next hype curve associated with newer technologies, such as social network mining.

However, as long as the world keeps producing data of all kinds (including text, web pages, images, sounds, etc.) at an ever increasing rate, the demand for data mining will continue to grow.

Table 7 Words with in 2006 KDnuggets news not present in 2005

Word06	Count In 06	Comment
Digitas	29	company which started to hire data miners
Nees	27	NEES consortium looking for Data curator
Geniq	25	software promotions
Choicestream	25	company which started to hire data miners
Visumap	23	software promotions
Allstate	21	company which started to hire data miners
Textanalyticsnews	20	new text analytics summit
Linguamatics	19	text mining company
Tenjinno	17	Tenjinno Machine Translation Competition
Predictiveanalyticstraining	15	course website
Predictionimpact	15	company offering predictive analytics training
Thomson	14	company which started to hire data miners
Subpoena	14	Google Defies US Subpoena story
MLSS	14	Machine Learning Summer School
Seti	13	SETI at home project
Quadrant	13	Gartner magic quadrant for Customer Data Mining

References

1. KDD-99 Panel Report , SIGKDD Explorations, 1999
2. KDnuggets News 99:n24, item2, E.Piphany to Buy RightPoint for $ 393 Mln, http://www.kdnuggets.com/news/99/n24/i2.html
3. KDnuggets News 2000:n01, item1, Vignette buys DataSage for $ 555 million, http://www.kdnuggets.com/news/2000/n01/i1.html
4. KDnuggets News 2000:n07, Exchange application acquires knowledge stream partners, www.kdnuggets.com/news/2000/n07/i1.html