Deriving Technology Indicators from Corporate Websites: A Comparative Assessment Using Patents

ETH zürich

KOF Swiss Economic Institute

Deriving Technology Indicators from Corporate Websites: A Comparative Assessment Using Patents

Sebastian Heinrich

KOF Working Papers, No. 512, July 2023

KOF

KOF

ETH Zurich
KOF Swiss Economic Institute
LEE G 116
Leonhardstrasse 21
8092 Zurich, Switzerland

Phone +41 44 632 42 39
www.kof.ethz.ch
kof@kof.ethz.ch

Deriving Technology Indicators from Corporate Websites: A Comparative Assessment Using Patents

Sebastian Heinrich
heinrich@kof.ethz.ch

Swiss Economic Institute
ETH Zurich
Leonhardstrasse 21
8092 Zurich
Switzerland

April 21, 2023

This paper investigates the potential of indicators derived from corporate websites to measure technology related concepts. Using artificial intelligence (AI) technology as a case in point, I construct a 24-year panel combining the texts of websites and patent portfolios for over 1,000 large companies. By identifying AI exposure with a comprehensive keyword set, I show that website and patent data are strongly related, suggesting that corporate websites constitute a promising data source to trace AI technologies.

Keywords: corporate website, patent portfolio, technology indicator, text data, artificial intelligence

JEL Codes: C81, O31, O33

Declarations of interest: none

1 Introduction

How can we measure the most recent technological developments in order to determine their economic implications? Throughout the initial phases of their life cycles, there is only limited data available to track technologies, with survey or patent data becoming widely available only after a significant delay (Raj and Seamans, 2018). Even later in its life cycle, a technology's development might be too dynamic to capture the most recent advancements using patent or survey data. Thus in rapidly developing fields such as artificial intelligence (AI), the lack of timely data does not only limit scientific research, but also thwarts informed policy responses. An additional important challenge regarding AI as a general-purpose technology is that its widespread usage by firms without related development activities is not captured by established data sources like patents. To bridge this gap in data availability, this paper proposes the use of web-based indicators and compares them with the more established patent data.

Using corporate websites, I develop text-based indicators to investigate technology related questions using AI as a case in point. I compile a firm level panel including 18 million web pages of 2,290 companies over 24 years, complementing the corporate website data with the company's patent portfolios, a widely established data source. As far as the author is aware, this is the first panel and analysis of its kind. This study evaluates this novel data using a comparative approach, contrasting the websites with the respective patent portfolios, using cross-sectional, time series and panel data. I find strong agreement between the novel web-based measures and the more established patent indicator, suggesting that corporate websites are indeed a valuable data source to track AI related developments.

2 Literature

This study contributes to AI research in economics, where recent advances have seen the use of various novel data sources. Webb (2019) and Acemoglu et al. (2020) use data on job descriptions and job postings to study the impact of AI on the labor market. Similarly, Babina et al. (2020) use employee resumes data to investigate the effect of AI on firm growth and product innovation. Finally, Baruffaldi et al. (2020) use open source software repositories to assess trends in AI related developments.

This paper also ties into the emerging literature on technology indicators derived from corporate websites. Héroux-Vaillancourt et al. (2020) validate website-based indicators with survey-based indicators, finding that website data can be used as a complement to direct measures or as substitute indicators for broad measures like importance of R&D. Kinne and Lenz (2021) are able to predict if firms are innovative based on their websites, using a deep learning approach trained on German survey data. Ashouri et al. (2022) develop a large scale data set containing web-based firm level indicators for the EU and UK.

Remarkably, there is hardly any work based on panel data, in spite of its strong analytical advantages. Most notably, Arora et al. (2020) use a short 4-year panel containing web page indicators for a few hundred companies to study strategic change in manufacturing. Similarly, Gök et al. (2015) compile 9-years of archived web data, derive text-based R&D activity indicators and compare them to other data sources, surprisingly not taking advantage of the panel structure. Therefore, this study appears to be first to establish and analyze a long-run panel of company websites.

3 Data and Method

The website panel I use for this study comprises 2,290 companies and covers 24 years (1996 to 2019), resulting in over 18 million web pages, downloaded during the autumn of 2020. The sample is based on the industrial R&D investment scoreboard provided by the European Commission, which records the largest R&D spenders globally. In each panel year, the latest version of every web page available in the archive was downloaded for the sample companies. To evaluate the potential of this novel data source, I merge the yearly website data with data from all patents filed by companies in my sample in the respective year. The patent data is obtained from Google Patents and Patstat, a database by the European Patent Office. I use patent families, a standard procedure to avoid double counting.

I apply a comprehensive keyword search to both data sources, using 38 terms related to AI technologies (Tables A.8 and A.9). Potential keywords are generated by extracting common noun phrases from the titles and abstracts of patents and scientific publications containing the term *artificial intelligence*, and are subsequently narrowed down using a publicly available AI glossary and manual inspection. The results of the search based on the final set of 38 keywords show that 2.49% of web pages and 0.24% of patents contain AI keywords, while over 30% of patent portfolios, and about 70% of websites belonging to companies in the sample contain documents that include AI keywords (Table A.1).

Methodologically, I follow a comparative approach, carving out the relative characteristics of the two data types by analyzing the (1) cross-section, (2) time series, and (3) panel. First, for the cross-section of n companies i, I compute ϕ, the Matthews correlation coefficient (MCC), between AI mentions in web pages and patents to uncover the extent to which the two data types agree on AI exposure at the company level and under what condition regarding their intensive margin. The MCC constitutes a special case of the Pearson correlation coefficient for where both variables are binary. Starting from AI document counts $D = \{ai_i^{web_c},\ ai_i^{pat_c}\}$ of website and patents, I create two binary measures $ai_i^{web_b}$ and $ai_i^{pat_b}$, using an indicator function $\mathbf{1}_{(x_i >= l)}$, where l is a threshold of minimally required AI documents $x \in D$ needed for company i to be determined as having AI exposure. The MMC can be calculated as follows:

$$\phi = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \tag{1}$$

where tp, tn, fp, and fn refer to the cells of a 2×2 confusion matrix, summing up the companies within each of the four comparison types resulting from the two binary variables, namely *true positive*, *true negative*, *false positive*, and *false negative.* However, the MMC is a correlation and does not require any ground truth, as the terminology might suggest.

Second, I construct time series of aggregated AI document counts for both data types to shed light on their temporal association. I calculate correlation coefficients between the web page count time series w and the patent count time series p^{shift}, shifted by s years along its time index. Computing the Pearson correlation coefficient ρ for multiple shifts s, I identify the $shift^{\rho_{max}}$ where ρ is highest:

$$shift^{\rho_{max}} = \arg\max_{s} corr(w,\ p^{shift}) \tag{2}$$

The results are provided in Section 4 and contrasted with the time lags between the first introduction of every keyword into each of the two data types.

Finally, I use linear panel regression models to further study the relation between patents and web pages at the company and sector level. I apply linear fixed-effects regression to analyze the relation between the two document types $D = \{ai^{web_{ln(c)}},\ ai^{pat_{ln(c)}}\}$, using their *log* transformed counts. The included fixed-effects control for unit specific characteristics that do not vary over time, allowing to estimate within unit effects (firm or sector). To explore the association between the two document types, I use dependent and independent variables interchangeably: $y \in D$ and $x \in D \setminus y$. Separate models are provided for either the company or the sector level and for a series of yearly shifts $s \in \{s \mid s >= -3,\ s <= 3\}$ applied to the independent variable:

$$y_{it} = b_0 + b_1 x_{it-s} + c_i + e_{it} \tag{3}$$

where i refers to the respective level, c_i are unit fixed effects, and e_{it} are i.i.d. normally distributed error terms.

4 Results and Discussion

This section presents the results for (1) the cross-sectional, (2) the time series and (3) the panel analysis. First, we inspect the MCC for a series of AI document thresholds l. Figure 1 shows that ϕ initially increases along the x-axis up to an AI web page threshold of $l \approx 30$, where correlations of up to 0.4 are achieved, while steadily decreasing for thresholds $l > 50$. The levels of the four curves, representing different AI patent thresholds, indicate that increasing patent thresholds slightly improve ϕ. The patent threshold peeks for very low l, followed by a steady decline, as Figure 2 shows. The initial rise of ϕ is only visible when requiring a modest number of AI web pages.

Figure 1: MCC for varying AI Web Page Thresholds

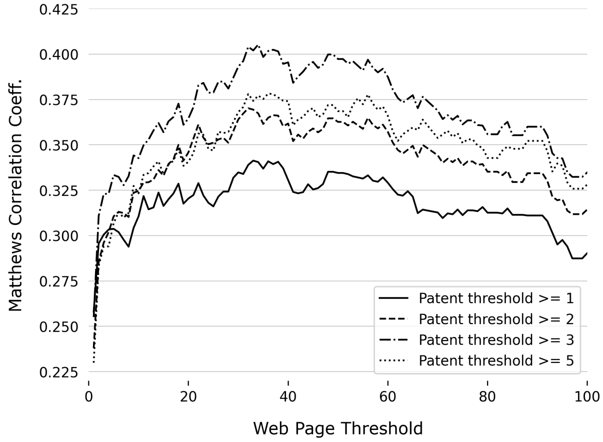
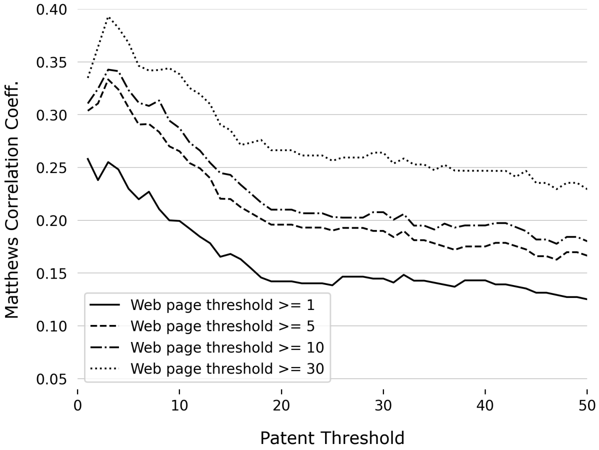


Figure 2: MCC for varying AI Patent Thresholds

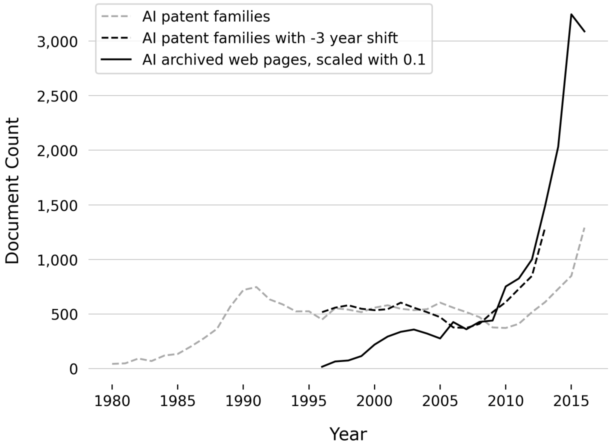


Matthews correlation coefficient between patent and websites for different configurations of thresholds. The threshold for one document type is varied while the other is kept constant for levels indicated in the legend.

Results show that a moderate raise of AI document thresholds benefits the correlation of web and patent indicators. However, by setting higher thresholds for the patent sample, the more inclusive website sample of AI companies moves closer to the more restrictive patent sample, possibly diminishing one of the key benefits of corporate websites, namely accounting for AI exposure beyond the development activity represented by patents.

Second, the shifted time series correlations are illustrated in Figure 3, where we see a considerable upward trend depicting significant growth in AI related activity. The correlation coefficient peaks at a shift of $s = -2$ years with $\rho = 0.94$, staying in a similar range for -3 and -4 year shifts with $\rho > 0.9$, where a -3 year shift seemingly depicts the strongest visual overlap. This means that web page volumes grow ahead of patent volumes. However, when taking into account the first appearance of a keyword on a web page and in a patent of a given company, new keywords tend to occur first in patents. This means that, even though websites precede patents in growth of aggregated volumes, the introduction of new knowledge commonly happens in patents.

Figure 3: Temporal Shifts of AI Technology Trajectories



AI trajectories for aggregated time series of corporate web pages and patents, including time shift.

Finally, the regression results are presented in Figures 4 and 5 as coefficient plots. The clearly visible inverted u-shapes depicting decreasing coefficients in both temporal directions show the strong temporal association of the two data types. At the company level, the relation appears to be mostly symmetrical around the zero time shift; however, the web page coefficients appear to be shifted marginally to the past, and the patent coefficients to the future. This might be taken to suggest that, on average, companies start to increase coverage of their respective technologies on their websites slightly prior to submitting their patent applications. Sector level results show a divergent pattern for shifts $s < 0$, where the association between past web pages and current patent counts remains high. This pattern supports the results of the shifted time series, where website volumes start to grow ahead of patent volumes.

Figure 4: Company Level Coefficients

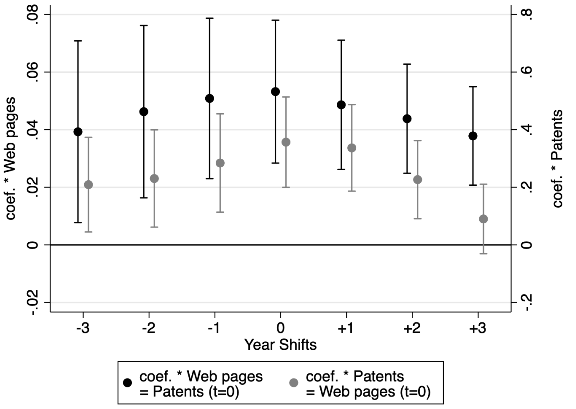
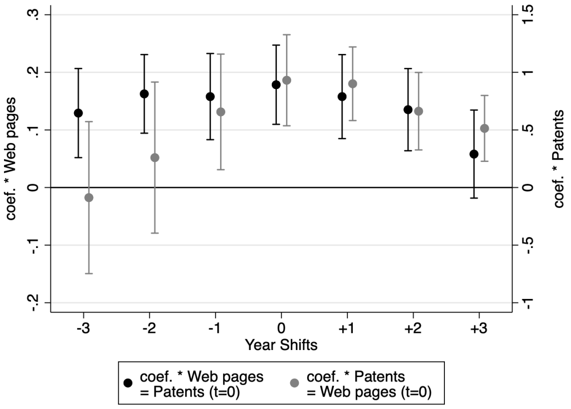


Figure 5: Sector Level Coefficients



Regression coefficients at the company and sector level with 95% confidence intervals. Results are provided for regressing patent counts (ln) on web page counts and vice versa. The legend indicates the model.

5 Conclusion

In this paper, I compile and evaluate a long-run panel derived from the publicly available websites of over 2,000 companies across 24 years, and complement the resulting 18 million corporate web pages with the respective companies' patent portfolios. The results of the subsequent keyword search confirm that technical keywords are prevalent on corporate websites, rendering this data source suitable to compile respective indicators.

At the cross-sectional level, I use the Matthews correlation coefficient to document the possibility of adjusting the novel web data towards the established patent data via document thresholds. The correlation between the binary indicators based on patents and websites increases strongly with higher web page thresholds. This result shows that the number of web pages containing technology keywords is of considerable importance when classifying firms into AI and non-AI firms according to web data.

The analysis of aggregate time series illustrates the well-known major increase in AI related activities taking place in recent years, especially regarding the key technological breakthroughs in deep learning technologies around 2010. Website and patent indicators are strongly associated, reaching a maximum correlation at a small temporal shift. While growth in web page volumes precedes that of patent volumes, patents clearly lead in the introduction of new knowledge. Thereby the comparison of aggregated time series confirms that corporate website data captures aggregate time trends and therefore appears promising for monitoring tasks.

Regression analysis at the company and sector levels confirm the relatedness of corporate websites and patent portfolios evident in the cross-sectional and time series analysis. At the company level regression results indicate that AI related activity measured by patent portfolios and websites are temporally synchronous. At the sector level results show a constantly strong relation of past web pages with current patents. This is in agreement with the finding from time-series analysis that the growth of aggregated web page counts slightly precedes that of aggregated patent counts.

Overall, the results document a strong relationship between the two data types, with marginally diverging patterns at different levels of aggregation. Besides specific caveats, historic data from web archives appear to provide a promising research opportunity, complementing established data such as patents. Indicators derived from corporate websites further appear to be well suited to monitor the development of a technologies and can thus potentially provide a basis to inform policy, especially throughout early or highly dynamic phases of technological development.

Data Availability Statement

Derived data supporting the findings of this study are available from the corresponding author S.H. on request. The web data that support the findings of this study are openly available from the Internet Archive at https://archive.org/. The author does not directly re-distribute any text of corporate websites. Patent data is openly available from Google Patents via the Google cloud platform. Patstat is subscription-based, available from the

European Patent Office for an annual fee.

Acknowledgements

I am grateful to Martina Jakob, Martin Wörter, Johannes Dahlke, conference participants at EPIP 2021, and anonymous reviewers for their invaluable comments.

References

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo (2020). *AI and Jobs: Evidence from Online Vacancies*. Working Paper 28257. National Bureau of Economic Research.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb (2019). *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Arora, Sanjay K, Yin Li, Jan Youtie, and Philip Shapira (2016). "Using the wayback machine to mine websites in the social sciences: a methodological resource". *Journal of the Association for Information Science and Technology* 67.8, pp. 1904–1915.

Arora, Sanjay K, Yin Li, Jan Youtie, and Philip Shapira (2020). "Measuring dynamic capabilities in new ventures: exploring strategic change in US green goods manufacturing using website data". *The Journal of Technology Transfer* 45.5, pp. 1451–1480.

Ashouri, Sajad, Arho Suominen, Arash Hajikhani, Lukas Pukelis, Torben Schubert, Serdar Türkeli, Cees Van Beers, and Scott Cunningham (2022). "Indicators on firm level innovation activities from web scraped data". *Data in brief* 42, p. 108246.

Babina, Tania, Anastassia Fedyk, Alex Xi He, and James Hodson (2020). "Artificial intelligence, firm growth, and industry concentration". *Working paper.*

Baruffaldi, Stefano, Brigitte van Beuzekom, Hélène Dernis, Dietmar Harhoff, Nandan Rao, David Rosenfeld, and Mariagrazia Squicciarini (2020). "Identifying and measuring developments in artificial intelligence". *OECD.*

Cockburn, Iain M., Rebecca Henderson, and Scott Stern (2018). "The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis". Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, pp. 115–146.

Gök, Abdullah, Alec Waterworth, and Philip Shapira (2015). "Use of web mining in studying innovation". *Scientometrics* 102.1, pp. 653–671.

Héroux-Vaillancourt, Mikaël, Catherine Beaudry, and Constant Rietsch (2020). "Using web content analysis to create innovation indicators—What do we really measure?" *Quantitative Science Studies* 1.4, pp. 1601–1637.

Kinne, Jan and David Lenz (2021). "Predicting innovative firms using web mining and deep learning". *PloS one* 16.4, e0249071.

Raj, Manav and Robert Seamans (2018). "Artificial Intelligence, Labor, Productivity, and the Need for Firm-Level Data". Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, pp. 553–565.

Webb, Michael (2019). "The impact of artificial intelligence on the labor market". *Available at SSRN 3482150.*

A Appendix

A.1 Data Acquisition Procedure

This section provides a detailed description of the acquisition process for domain names, web pages, and keywords:

1. *Domain names:* In order to acquire the domain names for companies in the sample, I use a semi-automated process. I apply this process to all unique company names. First, I search the company names in five different online services: Clearbit (company search engine), DuckDuckGo (search engine), Google Search (search engine), Wayback Machine (Internet Archive text search) and Wikidata (structured data equivalent to Wikipedia). The idea of using these services is to have access to recent information by search engines, as well as past information via archives and knowledge bases.

   After obtaining the top 5 search results from all services, I filter the domain names according to the agreement between the search services and the similarity of the domains with the company names, computing Levenshtein distances.[1] By extensive manual inspection, I derive the following four rules. Domains count as a match to company names (1) when either name and domain are exactly the same, (2) when at least 4 services agree on the domain, (3) when three services agree and the Levenshtein distance is at least 0.9 (range is 0 to 1), or (4) when at least 4 services agree and Levenshtein distance shows a perfect partial score.

2. *Web pages:* In each panel year the most recent version of every web page available in the Internet Archive belonging to one of the sample companies was downloaded. Thus, I used the maximum depth and breath of web pages per company available in the Internet Archive. The number of pages is not balanced over companies or years, as can be seen in Table A.2. Coverage of corporate web pages is determined by the

[1] For Wikidata I obtain the official website property.

procedures implemented by the Internet Archive. Web pages are retrieved by querying an index and the subsequent download of the actual document. Please refer to Arora et al. (2016) for a detailed documentation of the Internet Archive.

3. *Keywords:* In contrast to other studies where keywords are selected by experts (e.g., Cockburn et al., 2018; Baruffaldi et al., 2020), I follow a data driven procedure to select keywords describing AI technologies. I start with a search for the single keyword ”artificial intelligence” in scientific and patent documents. From each scientific publication or patent document returned by this search, I extract candidate keywords from their content.[2] The resulting candidates are subsequently validated against a glossary on artificial intelligence, selecting only terms contained in this glossary.[3] Finally, I exclude ambiguous terms manually, e.g., ”online learning”, referring to a machine learning technique, as well as a teaching method. The linkage of keywords to Wikipedia further allows to obtain synonyms and translations of the keywords, which I use in the subsequent search process.

   The validation step via a human curated glossary is beneficial to narrow down the scope of the candidate keyword set, which can be large. On the other hand, starting the search from the patent and scientific literature provides the keyword collection with credibility, as they originate from the relevant literature, where they occurred in the context of the term ”artificial intelligence”. In a final step all selected keywords, including the original keyword ”artificial intelligence”, are used to search patent documents and company web pages to obtain documents related to AI.

[2] This includes word n-grams and noun-phrases. N-grams are continuous sequences of n words. Noun-phrases are word sequences with a noun as head.

[3] I use a glossary from Wikipedia:https://en.wikipedia.org/wiki/Glossary_of_artificial_intelligence

A.2 Descriptive Statistics

Table A.1: Firm Level Data Sets and Keyword Search Results

Sample	Data	Type	Total Count	AI Count	AI %-share
Companies identified in Patstat	Patent Families	Company	1,800	542	30.11
		Document	9,276,764	21,954	0.24
Companies identified in Internet Archive	Web pages	Company	2,290	1,603	70.00
		Document	18,003,091	448,188	2.49
Companies identified in Patstat and Internet Archive	Patent Families	Company	1,042	364	34.93
		Document	5,481,470	15,819	0.29
	Web pages	Company	1,042	766	73.51
		Document	9,076,287	255,995	2.82

Overview of sample data sets. The table differentiates between availability of companies within the patent and web page data sets.

Table A.2: Distribution of Web Pages per Company

Measure	Web pages per company	AI web pages per company
Company count	2,290.00	1,603.00
mean	7,861.61	279.59
std	11,049.76	1,500.70
min	10.00	1.00
10%	304.70	1.00
20%	663.80	3.00
30%	1,252.40	6.00
40%	2,142.20	10.00
50%	3,497.00	18.00
60%	5,377.80	32.00
70%	8,144.40	67.40
80%	13,165.00	137.60
90%	21,632.90	415.20
max	129,259.00	36,989.00

Distribution of web pages per company, total and AI related. The table contains the company count, average pages, standard deviation, min, max, and percentiles.

Table A.3: Web Page Panel Descriptive Statistics

Year	Company Count	AI Company Count	AI Company %	Web page Count	AI Web page Count	AI Web page %
2019	2,232	1,075	48.16	1,300,561	120,046	9.23
2018	2,189	885	40.43	699,761	51,962	7.43
2017	2,196	898	40.89	1,377,677	57,010	4.14
2016	2,202	817	37.10	1,337,006	30,891	2.31
2015	2,166	720	33.24	1,411,965	32,455	2.30
2014	2,127	619	29.10	1,149,357	20,330	1.77
2013	2,116	573	27.08	1,223,302	14,902	1.22
2012	2,032	499	24.56	1,010,273	10,001	0.99
2011	2,041	447	21.90	867,883	8,254	0.95
2010	1,906	415	21.77	802,174	7,521	0.94
2009	1,887	363	19.24	686,212	4,402	0.64
2008	1,925	320	16.62	673,009	4,279	0.64
2007	1,854	328	17.69	630,418	3,602	0.57
2006	1,767	352	19.92	754,570	4,261	0.56
2005	1,708	248	14.52	411,911	2,760	0.67
2004	1,740	276	15.86	659,200	3,210	0.49
2003	1,671	310	18.55	696,561	3,577	0.51
2002	1,649	320	19.41	668,801	3,368	0.50
2001	1,641	323	19.68	560,048	2,932	0.52
2000	1,455	248	17.04	273,199	2,195	0.80
1999	1,241	185	14.91	153,556	1,147	0.75
1998	1,038	139	13.39	91,974	741	0.81
1997	667	126	18.89	107,716	651	0.60
1996	536	45	8.40	29,190	168	0.58

Yearly total and AI document and company counts for the website panel, including relative measures.

A.3 Regression Tables

Table A.4: Sector Level regression AI Web Pages on AI Patents

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Web page	Web page	Web page	Web page	Web page	Web page	Web page	Web page	Web page	Web page
	count, ln	count, ln	count, ln	count, ln	count, ln	count, ln	count, ln	count, ln	count, ln	count, ln
F3.Patent count,	0.251							0.0671		-0.136
ln	(1.72)							(0.57)		(-1.93)
F2.Patent count,		0.425*						0.228		0.213
ln		(2.45)						(1.91)		(1.95)
F.Patent count,			0.527***					0.0649		0.200
ln			(5.70)					(0.58)		(1.86)
Patent count, ln				0.580***				0.568	0.379*	0.414*
				(5.67)				(1.94)	(2.66)	(2.84)
L.Patent count,					0.442***				0.303*	0.0954
ln					(4.04)				(2.61)	(0.80)
L2.Patent count,						0.157			0.133	-0.0899
ln						(1.04)			(0.65)	(-0.42)
L3.Patent count,							-0.0846		-0.435	-0.205
ln							(-0.52)		(-1.99)	(-1.46)
L.Science count,	1.630***	1.711***	1.651***	1.754***	1.853***	1.893***	1.846***	1.562***	1.890***	1.627***
ln	(6.23)	(6.34)	(6.44)	(7.50)	(8.67)	(9.19)	(8.53)	(5.82)	(8.31)	(6.24)
Constant	-11.09***	-12.28***	-11.97***	-13.03***	-13.53***	-13.15***	-12.04***	-12.48***	-13.68***	-11.65***
	(-5.20)	(-5.74)	(-5.40)	(-6.55)	(-7.22)	(-6.85)	(-5.53)	(-6.25)	(-7.79)	(-4.89)
Observations	25693	27018	27150	27144	27018	26932	26547	20857	21619	17649

t statistics in parentheses
Note: Results corresponding to sector level coefficient plots in Figure 5. The regression contains sector fixed effects.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.5: Sector Level Regression AI Patents on AI Web Pages

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Patent	Patent	Patent	Patent	Patent	Patent	Patent	Patent	Patent	Patent
	count, ln	count, ln	count, ln	count, ln	count, ln	count, ln	count, ln	count, ln	count, ln	count, ln
F3.Web page	0.0732							0.0665		0.136
count, ln	(1.47)							(1.02)		(1.45)
F2.Web page		0.158**						-0.0719		-0.0911
count, ln		(3.24)						(-1.57)		(-1.57)
F.Web page			0.194***					0.133		0.138
count, ln			(4.90)					(1.64)		(1.78)
Web page count,				0.224***				0.100	0.282***	0.0901
ln				(6.83)				(1.92)	(4.89)	(1.82)
L.Web page					0.174***				-0.0985	-0.0653
count, ln					(5.33)				(-1.13)	(-0.88)
L2.Web page						0.192***			0.199**	0.202**
count, ln						(6.37)			(3.73)	(3.44)
L3.Web page							0.119**		-0.0292	-0.0380
count, ln							(2.95)		(-0.58)	(-1.04)
L.Science count,	-0.0587	-0.0817	-0.126	-0.163	-0.0611	-0.112	0.0377	-0.301	-0.287**	-0.533*
ln	(-0.38)	(-0.63)	(-1.30)	(-1.59)	(-0.82)	(-1.28)	(0.40)	(-1.76)	(-3.00)	(-2.59)
Constant	2.931*	2.732*	2.987**	3.213**	2.560**	3.016**	1.971*	4.415**	3.760***	5.909**
	(2.32)	(2.59)	(3.60)	(3.53)	(3.65)	(3.74)	(2.39)	(3.29)	(4.30)	(3.74)
Observations	29205	29609	28420	27144	25783	24432	23107	25599	22903	21468

t statistics in parentheses
Note: Results corresponding to sector level coefficient plots in Figure 5. The regression contains sector fixed effects.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.6: Company Level Regression AI Web Pages on AI Patents

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Web page count, ln	Web page count, ln	Web page count, ln	Web page count, ln	Web page count, ln	Web page count, ln	Web page count, ln	Web page count, ln	Web page count, ln	Web page count, ln
F3.Patent count, ln	0.0401							-0.125***		-0.126***
	(0.68)							(-3.77)		(-3.73)
F2.Patent count, ln		0.169*						0.0514*		0.0544*
		(2.53)						(2.02)		(2.18)
F.Patent count, ln			0.273***					0.171***		0.171***
			(3.63)					(4.66)		(4.75)
Patent count, ln				0.292***				0.195***	0.273***	0.196***
				(3.71)				(3.96)	(4.78)	(5.07)
L.Patent count, ln					0.221**				0.0366	0.0199
					(2.61)				(0.92)	(0.56)
L2.Patent count, ln						0.167*			-0.0219	-0.0362
						(2.00)			(-0.65)	(-1.11)
L3.Patent count, ln							0.144		0.00163	-0.00637
							(1.78)		(0.04)	(-0.15)
L.Science count, ln	0.132***	0.130***	0.127***	0.127***	0.129***	0.144***	0.162***	0.126***	0.159***	0.158***
	(17.03)	(16.89)	(16.70)	(16.61)	(16.71)	(16.76)	(16.72)	(16.52)	(16.40)	(16.37)
Constant	-0.976***	-0.962***	-0.946***	-0.944***	-0.955***	-1.092***	-1.258***	-0.938***	-1.236***	-1.230***
	(-14.52)	(-14.49)	(-14.39)	(-14.34)	(-14.40)	(-14.60)	(-14.76)	(-14.32)	(-14.62)	(-14.62)
Observations	38554	38554	38554	38554	38554	37512	36470	38554	36470	36470

t statistics in parentheses
Note: Results corresponding to company level coefficient plots in Figure 4. The regression contains company fixed effects.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.7: Company Level Regression AI Patents on AI Web Pages

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Patent count, ln	Patent count, ln	Patent count, ln	Patent count, ln	Patent count, ln	Patent count, ln	Patent count, ln	Patent count, ln	Patent count, ln	Patent count, ln
F3.Web page count, ln	0.0302***							0.00587		0.00482
	(3.54)							(1.21)		(1.01)
F2.Web page count, ln		0.0365***						0.0115*		0.0108*
		(3.82)						(2.57)		(2.44)
F.Web page count, ln			0.0410***					0.00895		0.00801
			(3.61)					(1.61)		(1.54)
Web page count, ln				0.0453***				0.0261**	0.0335***	0.0185**
				(3.60)				(3.04)	(4.38)	(3.22)
L.Web page count, ln					0.0435**				0.00935	0.00733
					(3.05)				(1.60)	(1.33)
L2.Web page count, ln						0.0386*			0.00802	0.00724
						(2.53)			(1.28)	(1.16)
L3.Web page count, ln							0.0312		-0.00177	-0.00225
							(1.94)		(-0.18)	(-0.23)
L.Science count, ln	0.0146***	0.0139***	0.0141***	0.0142***	0.0153***	0.0165***	0.0182***	0.0123***	0.0139***	0.0119***
	(5.70)	(5.37)	(5.44)	(5.46)	(5.77)	(5.72)	(5.68)	(4.72)	(4.39)	(3.66)
Constant	-0.0765***	-0.0711**	-0.0728**	-0.0734**	-0.0813***	-0.0908***	-0.104***	-0.0597**	-0.0708**	-0.0555*
	(-3.44)	(-3.18)	(-3.25)	(-3.27)	(-3.56)	(-3.60)	(-3.69)	(-2.69)	(-2.59)	(-2.02)
Observations	38554	38554	38554	38554	38554	37512	36470	38554	36470	36470

t statistics in parentheses
Note: Results corresponding to company level coefficient plots in Figure 4. The regression contains company fixed effects.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

A.4 Keyword Statistics

Table A.8: Web Pages

Keyword	Document Count	%-Share
robotics	138,644	23.15
artificial intelligence	108,027	18.04
machine learning	98,209	16.40
machine vision	75,563	12.62
predictive analytics	49,403	8.25
deep learning	37,710	6.30
speech recognition	25,882	4.32
neural network	16,197	2.70
computer vision	15,605	2.61
pattern recognition	10,514	1.76
natural language processing	10,506	1.75
expert system	3,179	0.53
artificial neural network	2,113	0.35
intelligent agent	1,911	0.32
self driving car	1,209	0.20
reinforcement learning	764	0.13
supervised machine learning	515	0.09
support vector machine	448	0.07
autonomous robot	384	0.06
inference engine	382	0.06
convolutional network	315	0.05
unsupervised learning	265	0.04
unsupervised machine learning	194	0.03
natural language generation	191	0.03
computational linguistics	189	0.03
statistical learning	84	0.01
computational intelligence	82	0.01
recurrent neural network	79	0.01
generative adversarial network	65	0.01
representation learning	49	0.01
reasoning engine	42	0.01
automated reasoning	42	0.01
artificial general intelligence	33	0.01
reasoning system	21	0.00
strong artificial intelligence	6	0.00
conversational agent	5	0.00
convolutional neural network	0	0.00
supervised learning	0	0.00

Table A.9: Patent Families

Keyword	Document Count	%-Share
speech recognition	6,585	27.62
neural network	5,515	23.13
machine learning	2,834	11.89
pattern recognition	2,804	11.76
expert system	923	3.87
natural language processing	837	3.51
artificial intelligence	695	2.91
inference engine	505	2.12
deep learning	479	2.01
convolutional neural network	327	1.37
artificial neural network	299	1.25
computer vision	277	1.16
machine vision	240	1.01
support vector machine	234	0.98
robotics	194	0.81
recurrent neural network	177	0.74
reinforcement learning	137	0.57
supervised learning	128	0.54
intelligent agent	95	0.40
autonomous robot	93	0.39
unsupervised learning	80	0.34
predictive analytics	68	0.29
reasoning system	63	0.26
statistical learning	47	0.20
supervised machine learning	37	0.16
convolutional network	31	0.13
reasoning engine	29	0.12
natural language generation	25	0.10
conversational agent	24	0.10
generative adversarial network	20	0.08
automated reasoning	13	0.05
unsupervised machine learning	13	0.05
self driving car	7	0.03
representation learning	5	0.02
computational intelligence	3	0.01
artificial general intelligence	0	0.00
computational linguistics	0	0.00
strong artificial intelligence	0	0.00