A List of Clean Tech Patents

Guan-Cheng Li
Fung Institute for Engineering Leadership, UC Berkeley

Kathryn Paisner
Director of Research and Business Development, IP Checkups, Inc.

Lee Fleming
Fung Institute for Engineering Leadership, UC Berkeley

College of Engineering
University of California, Berkeley

Fung Technical Report No. 2014.08.21
http://www.funginstitute.berkeley.edu/sites/default/files/Cleantech.pdf

August 21, 2014

The Coleman Fung Institute for Engineering Leadership, launched in January 2010, prepares engineers and scientists – from students to seasoned professionals – with the multidisciplinary skills to lead enterprises of all scales, in industry, government and the nonprofit sector.

Headquartered in UC Berkeley's College of Engineering , the Fung Institute combines leadership coursework in technology innovation and management with intensive study in an area of industry specialization. This integrated knowledge cultivates leaders who can make insightful decisions with the confidence that comes from a synthesized understanding of technological, marketplace and operational implications.

A List of Clean Tech Patents

Guan-Cheng Li
Coleman Fung Institute for Engineering Leadership, UC Berkeley

Kathryn Paisner
Director of Research and Business Development
IP Checkups, Inc.
1600 Shattuck Ave.
Berkeley, CA 94709

Lee Fleming
Coleman Fung Institute for Engineering Leadership, UC Berkeley

We provide a list of 26,399 clean tech patents. Each patent in the list is labeled as one of the six assignee types (Figure 1)- lone inventor, venture backed firm, small & medium firm, academic, government, and large firm- and as one of the six clean tech types- solar, wind, biofuel, geothermal, hydro, and nuclear.

7452425	Langhauser Associates, Inc.	Small & Medium Firms
7452461		Lone Inventor
7452465	United Utilities PLC	Venture Backed Firms
7452466		Lone Inventor
7452467	Andigen, LLC	Small & Medium Firms
7452702	Archer-Daniels-Midland Company	Large Firms
7452707	Danisco A/S, Genencor Division	Large Firms
7452710	E.I. du Pont de Nemours and Company	Large Firms
7452851	Afton Chemical Corporation	Large Firms
7452961	Sumitomo Chemical Company, Limited	Large Firms
7453023	Genencor International, Inc.	Large Firms
7453028	Stine Seed Farm, Inc.	Small & Medium Firms
7453029	Stine Seed Farm, Inc.	Small & Medium Firms
7453164	Polestar, Ltd.	Small & Medium Firms
7453165	Seadyne Energy Systems, LLC	Small & Medium Firms
7453166	Oceana Energy Company	Small & Medium Firms
7453167		Lone Inventor
7453168		Lone Inventor
7453187	Washington State University Research Foundation	Academic
7453242	Hitachi, Ltd.	Large Firms
7453972	Westinghouse Electric Co. LLC	Large Firms
7454968		Lone Inventor
7454990	Atlas Material Testing, LLC	Small & Medium Firms
7455502	SPAL Automotive S.R.L.	Small & Medium Firms
7455503	Pulse Group Holdings Limited	Small & Medium Firms
7455582		Lone Inventor

Figure 1. Screenshot of companies corresponding to certain patents labeled as one of the six assignee types

We performed the following steps to retrieve that list:

(1) Search the full text of US Patents from 1975-2012 for one or more of the following keywords: energy, power, combustion, turbine, petroleum, coal, solar, biofuel, and fuel.

We were left with 232,387 patents that met the search criteria.

(2) Obtain hand-coded training sets of U.S.-issued patents for each of our clean tech categories from IP Checkups, Inc.

For the solar, wind, biofuel, geothermal, and hydro categories, patent numbers were downloaded directly from the appropriate energy-related categories in IP Checkups' CleanTech PatentEdge™ online database. Keyword and patent class code searches were used to identify nuclear energy patents in U.S.P.T.O. patent data (via Thomson Reuters' Aureka® platform). All patents that fell into more than one category (*e.g.* a windmill with an integrated solar panel would be categorized as both “solar” and “wind”) were removed from the training sets; therefore, each training set comprised unique data.

The problem with using keywords to select patents is that we might have numerous false negatives. For example, a patent containing words such as photovoltaic, collector, heat, concentrator, and sunlight is highly likely a solar energy patent, even though it does not contain the keyword “solar” at all. Likewise, a patent having turbine, renewable, rotor, tower, blades, and windmill would be most likely talking about a power generator by means of wind. To minimize false negatives, we applied to a machine learning solution by using LIBLINEAR.

(3) Vectorize the patents by the unique terms used across them. We strip the patent texts of capitalization and revert all characters to lower case. Punctuation is also scrubbed. From here, the text is vectorized. We only consider single-word terms (no bigrams or trigrams). We treat each patent as a data point (or observation).

(4) Run LIBLINEAR using the training sets on the 232,387 patents identified for consideration. The top weighted terms for each energy category are:

Solar:

photovoltaic	0.1495	geothermal	0.0108
collector	0.1274	solar-powered	0.0106
heat	0.062	fresnel	0.0104
concentrator	0.0529	concentrated	0.0104
sunlight	0.043	dye-sensitized	0.009
electricity	0.0373	pond	0.0086
absorber	0.0335	reflective	0.0086
renewable	0.0329	heliostats	0.0083
reflector	0.0275	greenhouse	0.0082
wind	0.024	batteries	0.0081
concentrating	0.0229	insolation	0.0074
pv	0.0225	concentrators	0.0073
radiant	0.0205	boiler	0.0071
parabolic	0.0202	spacecraft	0.0069
collectors	0.0187	airship	0.0067
subcell	0.017	exchanger	0.0067
glazing	0.0158	multijunction	0.0066
rechargeable	0.0146	bandgap	0.0066
reflectors	0.0137	desalination	0.0061
rays	0.0131	photoactive	0.006
absorbing	0.0127	ev	0.0058
heliostat	0.0123	recharging	0.0058
thermal	0.0116	thermodynamic	0.0057
collecting	0.011	seawater	0.0056

Wind:

turbine	0.1958	winds	0.0092
renewable	0.0638	geothermal	0.0091
rotor	0.0614	propeller	0.009
solar	0.0519	hydroelectric	0.0089
tower	0.0485	harnessing	0.0087
electricity	0.0478	airfoils	0.0078
turbines	0.0446	vanes	0.0076
blades	0.0417	magnus	0.0076
windmill	0.0353	electrolyzer	0.0076
generator	0.0317	sodium-sulfur	0.0075
sail	0.0247	installations	0.0075
farm	0.0223	vertical-axis	0.0073
generators	0.0222	tidal	0.0072
nacelle	0.0194	air-jet	0.0069
kinetic	0.0156	microgrid	0.0066
grid	0.0153	park	0.0064
airfoil	0.0136	pylon	0.0064
ocean	0.0116	windward	0.0063
sails	0.0112	airship	0.0063
wind-powered	0.01	downwind	0.0063
rotors	0.0099	mast	0.0059
wind-driven	0.0096	propellers	0.0058
offshore	0.0094	meteorological	0.0058
aerodynamic	0.0093	airborne	0.0057

Biofuel:

biomass	0.0561	lumen-traveling	0.0172
algae	0.0502	depolymerizing	0.0168
ghg	0.0363	co-products	0.0164
pentose	0.033	tube-structured	0.0161
biodiesel	0.0324	bioreactors	0.0157
chara	0.0309	mediator-less	0.0156
bioproduct	0.0282	fermenter	0.0155
lignocellulosic	0.0266	ibr	0.0154
biofuels	0.0245	byproduct	0.0153
hexose	0.0226	parallel-operated	0.0153
greenhouse	0.0218	faeces	0.0151
algal	0.0212	cyanobacterium	0.0149
stillage	0.0202	scada	0.0148
triodia	0.0199	fermenters	0.0147
photosynthetic	0.0199	lipid-extracting	0.0147
photobioreactor	0.0198	photoautotroph	0.0146
biogasoline	0.0188	photo-bioreactor	0.0146
microalgae	0.0182	nano-fibrillated	0.0146
bioreaction	0.0178	btl	0.0144
multistory	0.0177	phototrophic	0.0143
renewable	0.0176	macroalgae	0.0143
chemoautotrophic	0.0175	exosporium	0.0143
butanol	0.0174	emissions	0.0142
bio-fuel	0.0173	switchgrass	0.0141
distillers	0.0172	chemosynthetic	0.014

Geothermal:

renewable	0.0909	thermodynamic	0.0269
brine	0.0833	fossil	0.0231
sodium-sulfur	0.0346	geo-thermal	0.0228
electricity	0.0324	aquifer	0.0219
subterranean	0.0311	underground	0.0215

steam	0.0208	solvent-enriched	0.0143
rankine	0.0198	heat-driven	0.0143
geothermally	0.0192	cavern	0.0142
uncondensible	0.019	ghe	0.0142
0048828125	0.0185	reinjection	0.014
open-hole	0.0185	tidal	0.0138
hydroelectric	0.0182	expander	0.0137
0046875	0.0181	hrsg	0.0137
rock	0.0179	non-condensible	0.0137
solute-bearing	0.0172	substream	0.0136
glhe	0.017	back-flushing	0.0136
noncondensable	0.0168	gte	0.0136
turbine-generator	0.016	campus	0.0136
turbine-motor	0.0153	sustainability	0.0133
co-generator	0.015	flashed	0.0132
brines	0.0149	aqueous-based	0.0129
exchangers	0.0148	61021000	0.0129
riparian	0.0145	acquifer	0.0129
electrolyzer	0.0144	carbon-neutral	0.0128
hydropower	0.0143	geologic	0.0127

Hydroelectricity:

tailwater	0.0225	water-phase-change	0.0198
water-storage	0.0219	aerosol-evaporation	0.0198
darrieus	0.0217	surface-cooling	0.0198
non-fossil	0.0217	vapor-cooling	0.0198
waterwheel	0.0215	air-currents	0.0197
water-supply	0.0212	thermally-powered	0.0197
cho-venturi	0.0203	boiling-bed	0.0197
waterwheels	0.0201	space-conditioning	0.0197
biomorphic	0.02	ferrel	0.0197
hydropower	0.02	vaporeforming	0.0197
tear-dropped	0.0199	sparsely-populated	0.0197
buoyantly	0.0198	afterbay	0.0197
a-shaft	0.0198	glaciers	0.0197
aerosol-evaporators	0.0198	solar-electric	0.0197
solar-energy-storage	0.0198	seaports	0.0197
exit-orifices	0.0198	boundless	0.0197
seawater-pumping	0.0198	caverns	0.0197
high-altitudes	0.0198	water-falling	0.0197
glacial-growth	0.0198	shelll	0.0197
simultan-eously	0.0198	duals	0.0196
potable-water-patent	0.0198	vast	0.0196
66881291	0.0198	cvhr	0.0196
glacial-melting	0.0198	ebullent	0.0196
cavern-water	0.0198	shorelines	0.0196
triple-cool	0.0198	low-altitude	0.0196

Nuclear:

neutron	0.1014	uranium	0.0242
neutrons	0.0645	mev	0.0241
gamma	0.0609	photons	0.0235
reactor	0.0568	scintillator	0.0232
fission	0.048	scintillation	0.023
nmr	0.038	fissile	0.0211
radioactive	0.0309	fissionable	0.0211
isotope	0.0265	gamma-ray	0.02
nuclei	0.0243	plutonium	0.02

tritium	0.0173	photomultiplier	0.01
isotopes	0.017	spins	0.0098
energies	0.0152	feedwater	0.0097
compton	0.0147	drywell	0.0095
moderator	0.0139	contractual	0.0093
deuterium	0.0137	fusionable	0.0093
actinide	0.0132	thermonuclear	0.0092
edge-on	0.0127	fertile	0.0089
decay	0.0126	transmutation	0.0088
frc	0.012	positron	0.0088
protons	0.011	deuterons	0.0088
renewable	0.0106	subatomic	0.0087
breeder	0.0104	nuclide	0.0083
spect	0.0102	radioelements	0.0083
department	0.0102	thorium	0.0082
wetwell	0.0101	osc	0.0081

(5) LIBLINEAR, on average, increased the clean tech sample size by a ratio of 1.33 and had a cross validation of 85.0135% to the original training set patents.

(6) We sent the newly classified patents back to IP Checkups, Inc. for manual review. IP Checkups categorized the solar, wind, and biofuel patents as “Relevant” or “Not Relevant” (Figure 2). We then removed the “Not Relevant” patents from the sample. Of the new patents identified by LIBLINEAR (*i.e.* the patents that were not in the original training set), 8% of the biofuels category and 24% of the solar and the wind categories were classified as “Relevant.”

The machine learning algorithm identified a number of patents that were not in the original data set (some of these patents had been intentionally removed, due to cross-categorization, as described above). In the Biofuels, Solar, and Wind categories, these new patents were manually sorted into two sub-categories, "Relevant" and "Non-Relevant." Within each "Non-Relevant" category (in each of the three technology subsets), erroneous results were further categorized by technology area (Figure 3). The idea was that each of these technology areas could be used as a "not" training set for the algorithm, so that it could do a better job of excluding non-relevant results in the future.

Results were identified as "Relevant" or "Non-Relevant" on the basis of the patents' claimed invention(s).

In general, the threshold for relevancy was rather low. Auxiliary technologies were classified as "Relevant," even if they could not be used to harvest energy on their own. For example, coatings for wind turbine blades and transformers used to connect a solar panel to the grid were both classified as "Relevant," but a long-chain fatty acid used as a lubricant (with no mention of biodiesel) would be considered "Non-Relevant." Technologies that could be used for energy harvesting and other purposes (e.g. a reflector used in LEDs and solar panels) were also classified as "Relevant."

Conversely, technologies that mentioned renewable energy sources yet were primarily directed at other applications were classified as "Non-Relevant." A patent describing a watch that could use a solar panel as a power source would fall into this category.

(7) We then manually reviewed the sample to assign assignee types, clean out dubious cases, correct countries (DE as a suffix can mean Deutschland or Delaware; CA as a suffix can mean Canada or California), and clean errors from original USPTO data.

The complete list can be downloaded at:

http://funglab.berkeley.edu/guanchengli/cleantech_patents.tsv

The list of patents has also been geographically visualized at:

http://funglab.berkeley.edu/cleantechx/

as can be seen in figure 2 and 3.

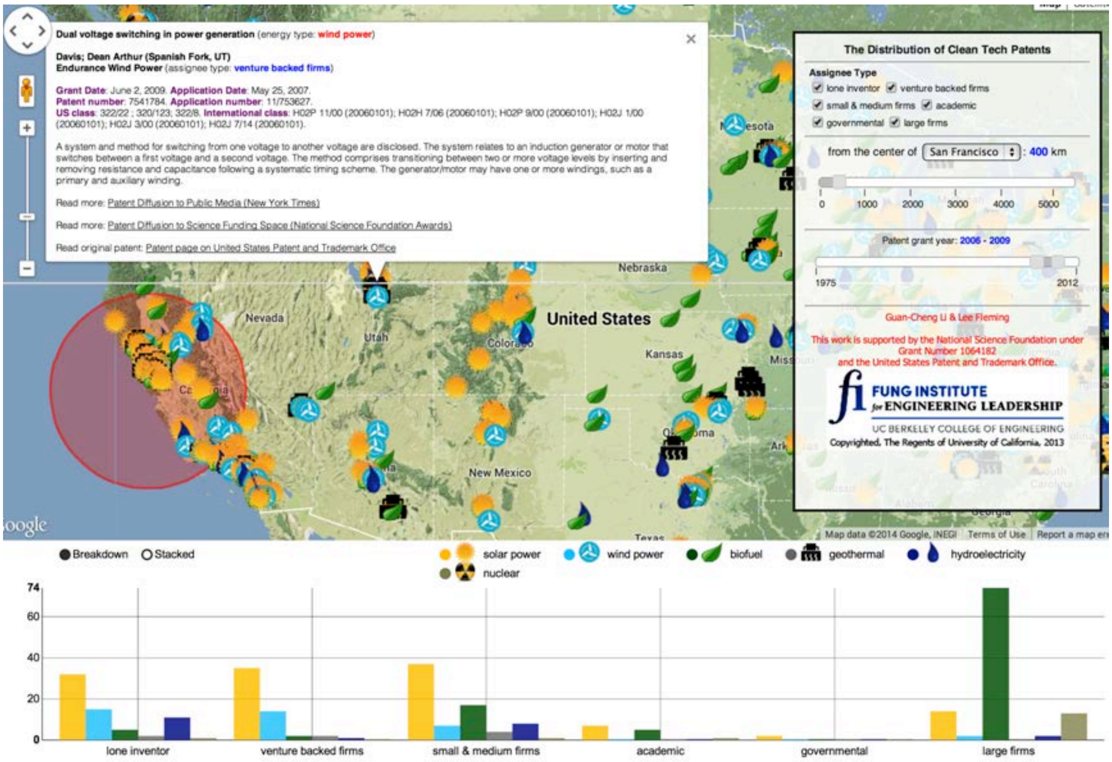

Figure 2. Geographical representation of patent list

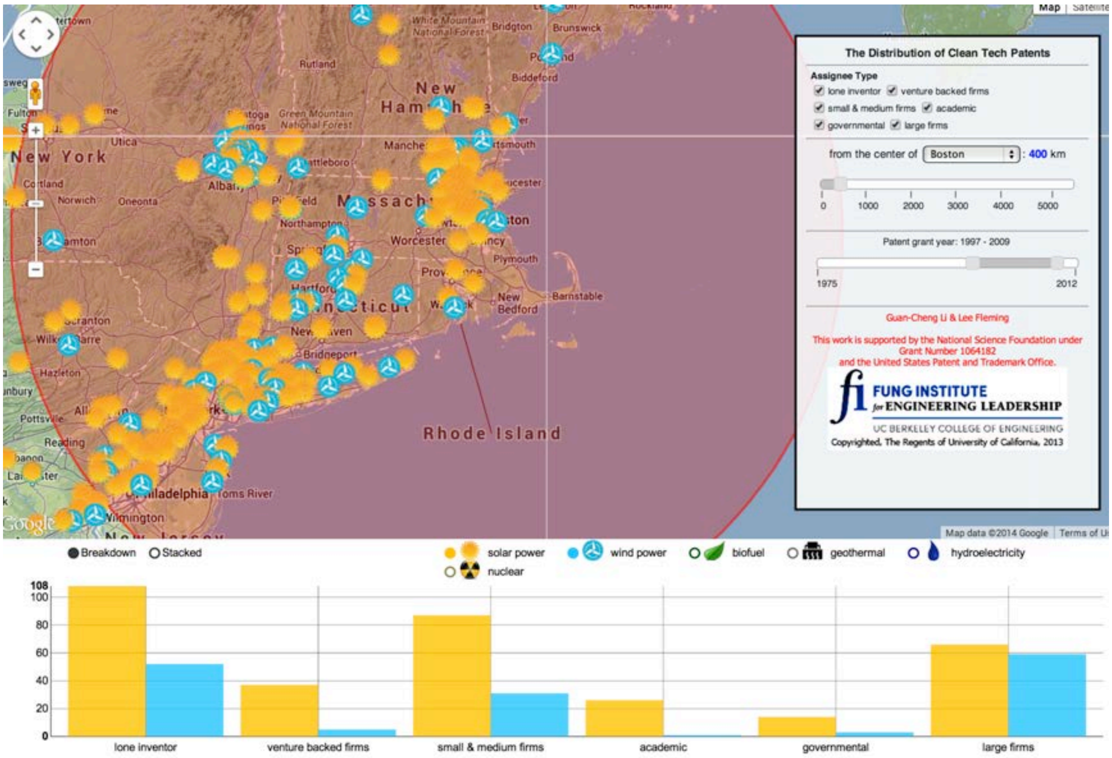

Figure 3. Geographical representation of patent list

References

1. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. "LIBLINEAR: A Library for Large Linear Classification", Journal of Machine Learning Research 9(2008), 1871-1874.

2. PatentCam (TM), a product of IP Checkups Inc,
https://www.patentcam.com/ ,
http://www.ipcheckups.com/wp-content/uploads/2011/06/PatentCAM.pdf .