



Innovative Green SMEs in China

– Using Multiple Data Source

Jie Ren

School of Management and Economics

Beijing Institute of Technology

Content



- ❖ **Introduction**
- ❖ **Background**
- ❖ **Governmental structure**
- ❖ **Data Base**
- ❖ **Result & Discussion**
- ❖ **Limitation**

Introduction



- Key Questions:
 - What is the current situation of China's Green SMEs?
 - What factors will influence the sustainable field in China: policy, location and global market
 - Unstructured data scraping and Chinese language analysis
 - Cooperation between Government, University and the Enterprises in China's Green Good Sector.
 - How do the Chinese green enterprises manage their international market?
- Challenges
 - None Structured data collection
 - Data integration (Different data source: Structured data + unstructured data; Chinese version company web pages + English version company web pages + Alibaba information)

Background



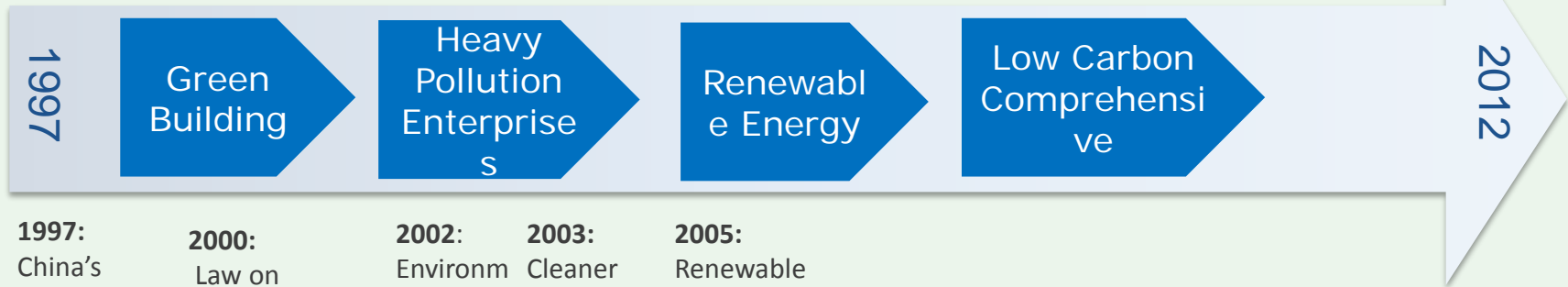
Projects

2006:
Top-1000 Energy-Consuming Enterprises Program (2006-2010)

2007:
Coal Bed Methane Mining Subsidy

2008:
-Biomass Power Generation Subsidy
-Wind Turbine Subsidy

2009:
-Solar Roof Project
-Subsidy on PHEV, EV and FC
-RMB387 billion on Smart Grid Project



Law

1997:
China's Energy Conservation Law

2000:
Law on Prevention and Control of Atmospheric Pollution" (Revised)

2002:
Environmental Impact Assessment Law

2003:
Cleaner Production Promotion Law

2005:
Renewable energy law

2005:
Medium and Long Term Plan for the Oil Refining Industry (2005-2020)

2006:
-11th 5-Year Plan
-Medium and Long Term Plan for Renewable Energy Development (2006-2020)

2007:
-National Program in Response to Climate Change
-Medium and Long Term Plan for Renewable Energy Development

2011:
12th 5-Year Plan

Plan and Programs

2000:
Medium and Long Term Science and Technology Development Plan(2000-2020)

Green Definition



General	[1]
Environmental	[2]General
	[3]Biological treatment
	[4]Air Pollution
	[5]Environmental Monitoring, Instrumentation and Analysis
	[6]Marine pollution control
	[7]Noise & Vibration control
	[8]Contaminated land reclamation & remediation
	[9]Waste management
	[10]Water supply and waste water treatment
	[11]Recovery and recycling
Renewable energy	[12]General
	[13]Wave & tidal
	[14]Biomass
	[15]Wind
	[16]Geothermal
	[17]Photovoltaic /solar

Emerging low carbon	[18]General
	[19]Alternative fuel vehicle
	[20]Alternative fuels
	[21]Electrochemical processes
	[22]Battery
	[23]Additional energy sources
	[24]Carbon capture & storage
	[25]Energy management
	[26]Building technologies

Data Selection



[Home](#) > [Advanced Search](#)

Advanced Search

[Products](#) **[Suppliers](#)** [Buying Requests](#)

Please input a keyword [Tips](#)

All words, any order Exact match Any words, any order

Refine by category:

Environment

Business type:

Manufacturer

Refine by country:

China (Mainland)

Gold Suppliers

OR [Post Buying Request](#)



Began in 1999, Alibaba.com is a B2B portal to connect Chinese manufactures with overseas buyers. It serves to bring together importers and exporters from more than 240 countries and regions.

Advantages

- More than 61 million registered members
- Began in 1999, 14 years history
- Active Enterprises in Global Markets
- Semi-structured Company information
- Detailed Company Profile
- Information of the product on shelves
- Special categories related with green technology for suppliers, as "Energy", "Environment"...

Disadvantages

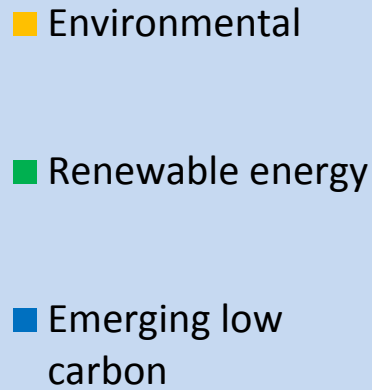
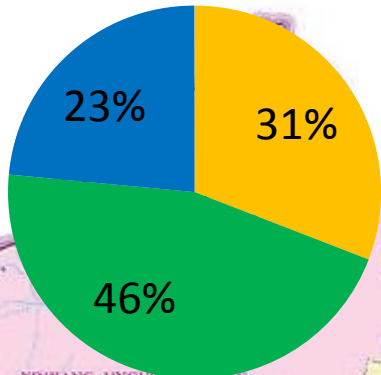
- Up to 50 characters in the searching textbox
- Boolean logic inadaptable
- Wildcard inadaptable
- Hard to filter, more manual work
- Most of the companies are cross sector in Alibaba categories

Data Description

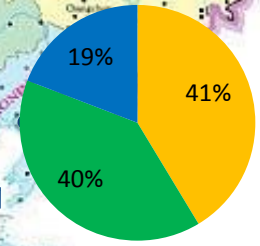


- 300 SMEs in Green Sector
 - 300 Alibaba company profiles
 - 191 Chinese version sites (63.67%)
 - 194 English version sites (64.67%)
 - 2004-2011 Wayback archive
- Established Year: 2002-2011
- Data Base
 - Company basic information from **Alibaba**
 - **Publication** from web of Science
 - 4 companies, 6 articles, 5 green papers
 - **Patent data** from Derwent
 - 58 companies, 582 patents, 364(62.5%) are green patents, 160 (43.9%) Solar field patents
 - **Company Webpages**
 - 57919 Chinese web pages
 - 42974 English web pages

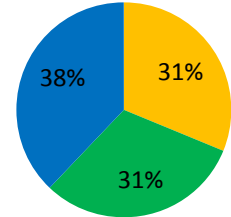
General



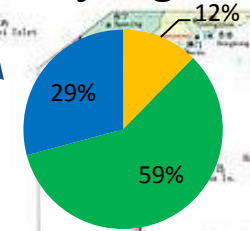
Shandong



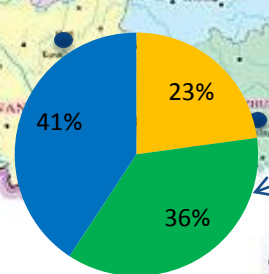
Jiangsu



Zhejiang



Guangdong



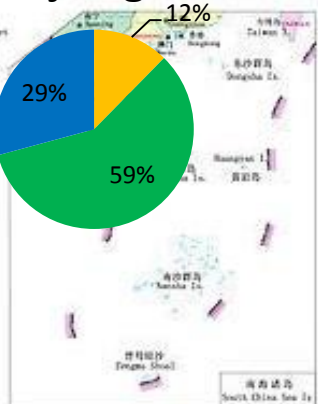
47

44

49

66

- Capital
- Provincial administrative centers
- Administrative center of prefectural cities
- Administrative centers of county cities
- County administrative centers





- Why “webscraping”?
 - Lack of data information for SMEs in China
 - Public data as a potentially source
 - Rich information in product
- Problems
 - How can website data be used?
 - How to integrate the data?

Administration Console for IBM Content Analytics with Enterprise Search - Mozilla Firefox

Administration Console for IBM Content Analytics with Enterprise Search

Search Customer | Analytics Customer | Log Out | Help | About

Collections > @raphene.2010 > Angstrom v2 : Web crawl space

Web Crawl Space

Learn more

Specify the URLs where the crawler is to start crawling. From these URLs, the crawler follows links to reach other pages, according to the crawling rules (click **Next** to specify pattern matching rules for including and excluding URLs). The list of domains to be crawled includes the start URLs.

* Start URLs
Type one URL per line, include the protocol, such as http://, and do not specify wildcard characters. If you enabled support for IPv6 addresses, enclose the URLs in brackets. For example: http://[2001:db8:0:1000::1]
http://web.archive.org/web/20100127005747/http://www.angstrommaterials.com/

Configure options for the entire crawl space

- Specify the keystroke to use for SSL connections
- Select the types of documents to include
- Specify the file extensions to exclude
- Specify the URL path depth
- Specify the language and code page
- If-ETag document-level security

Configure options for specific Web sites

- Crawl Web sites that are password-protected by HTTP basic authentication
- Crawl Web sites that are password-protected by NTLM authentication
- Crawl Web sites that are password-protected by form-based authentication
- Crawl Web sites that are served by HTTP proxy servers
- Configure rules for handling soft error pages
- Configure schedules for crawling specific Web servers

Back Next Finish Cancel

Administration Console for IBM Content Analytics with Enterprise Search - Mozilla Firefox

Administration Console for IBM Content Analytics with Enterprise Search

Search Customer | Analytics Customer | Log Out | Help | About

Collections > @raphene.2010 > Angstrom v2 : Web crawl space

Rules to Crawl HTTP Prefixes

Learn more

Specify the HTTP and HTTPS prefixes that you want to allow or forbid the crawler to crawl. The wildcard character (*) can occur one or more times in the URL. Examples: allow prefix http://*.ibm.com/public/* (crawls pages in the public directory on this domain) forbid prefix http://*.ibm.com/* (excludes all other directories on this domain)

Important: The order of the rules is significant. The crawler applies the first rule that matches a candidate URL.

Domains HTTP Prefixes IP Addresses

HTTP prefixes (type one prefix rule per line)

```
allow prefix http://web.archive.org/web/2010*angstrommaterials.com*
forbid prefix http://web.archive.org/web/*
forbid prefix http://*/cgi-bin*
forbid prefix http://*/.nsf/*?deletedocument*
forbid prefix http://*/.nsf/*deletedocument*
forbid prefix http://*/.nsf/*savedocument*
forbid prefix http://*/.nsf/*printable*
forbid prefix http://*/.nsf/collaps*
forbid prefix http://*/.nsf/frag*
forbid prefix http://*/.nsf/date*
```

Test specific URLs
URLs to test (type one URL per line and specify the protocol, such as http://)

IBM Content Analytics with Enterprise Search

IBM Content Analytics with Enterprise Search

rulebased:"US Companies" "ISE Corp"

Search Clear

Advanced Search Query Tree

142/1223 results matched

Facet Navigation

Document Cluster

- Terms of Interest
- Crawl Date
- Technology_Focus
 - environmental
 - renewable_energy
 - emerging_low_carbon
 - general
- Demos
- Manufacturing_Intensity
- Products
- Investors
- Venture_Capital
- Contextual Views
- Contextual views to analyze:
- Contextual view to search:

Search

Keywords	Frequency	Correlation
Battery	105	1.5
Lithium Ion	104	11.2
fuel cell	66	9.2
hybrid drive	64	11.4
hydrogen	29	3.5
battery	24	1.3
Fuel Cell	23	1.3
Hydrogen	23	1.7
Hybrid Drive	20	9.6
batteries	17	1.8
hybrid vehicles	16	8.3
alternative fuel	10	6.1
electric motors	10	5.1
charge	9	0.6
build	8	0.3
CO2	7	1.0
hybrid vehicle	7	4.6
electric vehicles	6	1.3
construction	6	0.2
Build	6	1.9
Hybrid Vehicle	4	3.3

Documents Facets Time Series Deviations Trends Facet Pairs Connections Dashboard

1157/1157 results matched

Facet 1 shows Subfacets Truncated Correlation

Facet 2 shows Keywords Highlight mode

Filter

- Part of Speech
- Phrase Constituent
- Named entity
- Person
- Location
- Organization
- My Keywords
- Document Cluster
- Terms of Interest
- Entity
- Predicate
- Crawl Date
- US Companies
 - Kompletech USA, Inc
 - Radiant Industrial Solutions
 - EnviroTech Systems
 - Akanis Energy
 - Sunika
 - MegTec
 - Genation
 - Asia Pacific Fuel Cell Technologies
 - ISE Corp

Search type: Subfacet search

Facet Path: /US Companies"

Value

Key

- US Companies
- Location
- Frequency
- Correlation Amount

Network diagram showing connections between entities like USA, UK, Germany, Australia, Japan, United States, etc.

Web variables



- number of pages
- Instances of Variables
- Demo
- products
- government links
- trials
- manufacturing Intensity
- R&D Instant
- university link
- venture capital
- greenness

E.g. Trials (Keywords: experimental, experimentation, exploratory, pilot, prelim, preliminary, provisional, tentative, test, testing, trial, in-process)



	Number of compan y	cover rate %	Instanc es	Instances per company	mean	min	max	media n
products	187	96.39	22907	122.50	1.019828	0.0142 98	4.7031 91	0.7297 3
trials	171	88.14	107569	629.06	0.505924	0.0043 33	2.0312 5	0.3037 97
government links	138	71.13	3052	22.11	0.12394	0.0004 33	1.0312 5	0.0687 29
manufacturing Intensity	181	93.30	17499	96.68	0.458349	0.0131 58	4	0.2426 41
university link	71	36.60	882	12.42	0.074704	0.0022 93	1	0.025
venture capital	44	22.68	211	4.79	0.024695	0.0013	0.1282 05	0.0151 59
greenness	117	60.31	4794	40.97	0.186849	0.0020 64	2.0681 82	0.0833 33
R&D Intensity	164	84.54	4615	28.14	0.302459	0.0031 1	2.0689 66	0.1418 04
Demo	147	75.77	4501	30.62	0.224178	0.0008 67	2	0.0909 09

Limitation, Challenges and future work



- Limitation of the work
 - Alibaba data base
 - Webscraping information missing
- Challenges
 - Non-structured website data integration
 - Chinese textual analysis
- Future work
 - Chinese web information analysis
 - Chinese and English version webpages comparasion
 - Wayback information analysis
 - Case study on special SMEs in China



Thank you!

Jie Ren
renjie_3@163.com