In [1]:
```
# import jieba
import pandas as pd
import numpy as np
from pymongo import MongoClient
%matplotlib inline
import matplotlib.pyplot as plt
```

從資料庫拉資料

In [2]:
```
client = MongoClient("mongodb://localhost:27017/")
db = client.news_textmining
news_title_collection = db.news_title_collection
news_amount = db.news_amount

cursor = news_title_collection.find()
dataset_main = pd.DataFrame([d for d in cursor])

cursor = news_amount.find()
news_amount = pd.DataFrame([d for d in cursor])
```

資料整理

對新聞標題作處理，將標號（e.g. 1.）去除
將日期拆開成年、月、日

In [3]:
```
dataset = dataset_main.drop(columns=["_id"])
dataset["title"] = dataset["title"].str.replace("(\d+)(\.)", "", regex=True)
dataset["title"] = dataset["title"].dropna()
dataset[dataset["title"].fillna("NA").str.contains("歐巴馬")]

news_amount = news_amount.drop(columns=["_id", "start_date", "key_word"])
news_amount["month"] = pd.to_numeric(news_amount["end_date"].str.split("-", expand = True)[1], errors = "coerce")
news_amount["year"] = pd.to_numeric(news_amount["end_date"].str.split("-", expand = True)[0], errors = "coerce")
news_amount = news_amount.drop(columns=["end_date"])
```

In [9]:
```
dataset
```

Out[9]:

	date	key_word	paper	position	position_code	reporter	title	month	year
3296	1990-02-07	兩黨分裂	聯合報	聯合副刊	29	京夫子	毛澤東和他的女人們	2.0	1990.0
3295	2001-03-21	兩黨分裂	聯合報	民意論壇	15	胡祖慶／東海大學政治系副教授（台中市	泛國民黨勢力 重組或解組？	3.0	2001.0
3294	2004-10-10	兩黨分裂	聯合報	雲嘉南新聞	B2	記者唐秀麗／嘉義報導	陳志周如願 棺木覆蓋黨旗	10.0	2004.0
3293	2012-09-06	兩黨分裂	聯合報	要聞	A4	蘇起	台灣民主的反思	9.0	2012.0
3291	2016-01-17	兩黨分裂	聯合晚報	2016大選民進黨布局	A7	記者徐偉真／台北報導	學者：周子瑜事件讓大黨吸回小黨票	1.0	2016.0
3645	1971-03-16	兩黨對立	聯合報		02	本報駐美記者 施克敏	傅爾布萊特與季辛吉之爭	3.0	1971.0
3644	1978-12-10	兩黨對立	聯合報	第二版	02	馮滬祥	從海外看國內選舉	12.0	1978.0
3643	1979-11-26	兩黨對立	聯合報		02	張忠棟	恢宏開放·積極進取	11.0	1979.0
3642	1983-11-18	兩黨對立	聯合報	綜藝 萬象	12	胡慕遠	笑談美國總統選舉	11.0	1983.0
3635	1988-11-10	兩黨對立	聯合晚報	話題新聞	02	記者尹復生／台北報導	五大法案 訴求平行線?	11.0	1988.0
3636	1988-11-10	兩黨對立	聯合晚報	話題新聞	02	記者胡遜/特稿	國防事務 何妨以'平常'心視之	11.0	1988.0
3637	1988-04-26	兩黨對立	聯合報	輿情	09	方明財(屏東市)	早餐的省思	4.0	1988.0
3638	1988-04-16	兩黨對立	聯合報	國內要聞	02	社論	朝野雙方具共同責任營建一個良性的溝通環境	4.0	1988.0
3639	1988-04-12	兩黨對立	聯合報	國內要聞	02	本報專欄組記錄整理	政黨政治往何處去?	4.0	1988.0
3640	1988-04-08	兩黨對立	經濟日報	要聞	02	台北訊	為總預算案激烈衝突	4.0	1988.0
3641	1988-01-28	兩黨對立	聯合報	••	03	聯合報系美加新聞中心策劃	台灣的應變與瞻望	1.0	1988.0
3628	1989-12-06	兩黨對立	聯合報	選舉糾紛	03	陳其南	選戰後朝野政黨政治的重新定位	12.0	1989.0
3629	1989-11-30	兩黨對立	聯合報	選舉新聞/話題	06	陳其南	台灣民主政治的前景與台獨問題	11.0	1989.0
3630	1989-10-09	兩黨對立	聯合報	社會觀察	04	記者胡文輝/特稿	立院便宜行事 自毀形象!	10.0	1989.0
3631	1989-09-29	兩黨對立	聯合晚報	話題新聞	03	記者葉陸裔／台北報導	'澄社'楊國樞：取消黑名單 有利選舉	9.0	1989.0
3632	1989-03-24	兩黨對立	聯合晚報	話題新聞	02	圖與文：邱勝旺	立法院會人數難湊	3.0	1989.0
3633	1989-03-01	兩黨對立	聯合報	焦點新聞	02	記者胡文輝／特稿	彈性節制緩解抗爭 兩黨協商良性發展	3.0	1989.0
3634	1989-01-14	兩黨對立	聯合晚報	要聞	01	記者尹復生/台北報導	不和就戰 三大法案談判底線 執政黨決縮小讓幅	1.0	1989.0
3615	1990-12-28	兩黨對立	經濟日報	要聞	02	本報訊	股市封關後的省思	12.0	1990.0
3616	1990-12-21	兩黨對立	聯合報	焦點新聞	02	記鄧蔚偉／台北報導	立院表決 確定公務員三法議事錄	12.0	1990.0
3617	1990-12-14	兩黨對立	聯合晚報	話題新聞	02	記者陳素秋／台北報導	心中有法案兩黨未放棄協調	12.0	1990.0
3618	1990-12-13	兩黨對立	聯合晚報	要聞	01	記者陳素秋/台北報導	黃華案 再使兩黨對立昇高	12.0	1990.0
3619	1990-12-06	兩黨對立	聯合晚報	台北市	10	記者曾淑玲/台北報導	周伯倫拒絕指名道姓 鬧成一堆	12.0	1990.0
3620	1990-11-20	兩黨對立	聯合報	經濟新聞	09	記者林美玲／台北報導	印花稅存廢 引發朝野兩黨對立	11.0	1990.0
3621	1990-11-17	兩黨對立	聯合晚報	要聞	01	記者陳素秋/台北報導	主權獨立之爭 似可轉圜	11.0	1990.0
...	...	...	...	...	...	...	...	...	...
4586	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4755	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5105	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5106	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5107	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5111	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5112	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5113	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5115	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5117	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

	date	key_word	paper	position	position_code	reporter	title	month	year
5123	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5124	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5128	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5131	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5162	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5177	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5193	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5200	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5201	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5203	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5205	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5206	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5207	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5216	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5217	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5224	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5229	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5231	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5245	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5263	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

5301 rows × 9 columns

各關鍵字的出現次數

比較剔除國際相關關鍵字後，新聞數量的增減。

In [4]:
```
tmp1 = dataset[["key_word", "title"]].groupby(by="key_word").count().sort_values(by="title")

dataset[dataset["position"]=="國際"] = np.nan
dataset[dataset["position"]=="國際焦點"] = np.nan
dataset[dataset["position"]=="國際財經"] = np.nan
dataset[dataset["position"]=="國際村"] = np.nan
dataset[dataset["position"]=="國際·運動"] = np.nan
tmp2 = dataset[["key_word", "title"]].groupby(by="key_word").count().sort_values(by="title")

tmp = pd.merge(tmp1, tmp2, on="key_word")
tmp
```

Out[4]:

key_word	title_x	title_y
政黨空轉	1	1
朝野分裂	4	4
兩黨分裂	6	5
國會分裂	19	11
國會惡鬥	19	15
國會停擺	20	20
國會對立	34	23
藍綠分裂	37	37
國會空轉	47	45
政黨衝突	48	48
國會衝突	59	56
兩黨衝突	84	81
政黨分裂	103	88
藍綠衝突	119	119
兩黨對立	130	114
兩黨惡鬥	135	117
朝野惡鬥	282	273
朝野衝突	348	343
政黨對立	458	444
藍綠惡鬥	608	608
藍綠對立	693	691
政黨惡鬥	991	970
朝野對立	1056	1032

資料對接

篩選政黨惡鬥、朝野對立、藍綠對立、超越藍綠關鍵字

In [5]:
```
dataset["month"] = pd.to_numeric(dataset["date"].str.split("-", expand = True)[1], errors = "coerce")
dataset["year"] = pd.to_numeric(dataset["date"].str.split("-", expand = True)[0], errors = "coerce")

dataset = dataset.sort_values(by=["key_word","year"])
des1 = dataset[dataset["key_word"] == "政黨惡鬥"][["year", "month", "key_word"]].groupby(by=["year", "month"]).count().
reset_index()
des1 = des1.rename(columns={"key_word":"政黨惡鬥"})
des2 = dataset[dataset["key_word"] == "朝野對立"][["year", "month", "key_word"]].groupby(by=["year", "month"]).count().
reset_index()
des2 = des2.rename(columns={"key_word":"朝野對立"})
des3 = dataset[dataset["key_word"] == "藍綠對立"][["year", "month", "key_word"]].groupby(by=["year", "month"]).count().
reset_index()
des3 = des3.rename(columns={"key_word":"藍綠對立"})
des4 = dataset[dataset["key_word"] == "藍綠惡鬥"][["year", "month", "key_word"]].groupby(by=["year", "month"]).count().
reset_index()
des4 = des4.rename(columns={"key_word":"藍綠惡鬥"})
des5 = dataset[dataset["key_word"] == "超越藍綠"][["year", "month", "key_word"]].groupby(by=["year", "month"]).count().
reset_index()
des5 = des5.rename(columns={"key_word":"超越藍綠"})

merged = pd.merge(des1, des2, on=["year", "month"], how="outer")
merged = pd.merge(merged, des3, on=["year", "month"], how="outer")
merged = pd.merge(merged, des4, on=["year", "month"], how="outer")
merged = pd.merge(merged, des5, on=["year", "month"], how="outer")

merged = pd.merge(merged, news_amount, on=["year", "month"], how="outer")
merged = merged.sort_values(by=["year", "month"])

merged = merged.loc[merged["year"]>=1990,:]
merged.fillna(0, inplace=True)
```

建構比率資料集（ratio）

In [6]:
```
ratio_month = pd.DataFrame({"year":merged["year"].unique().tolist()})
for kw in ["政黨惡鬥", "朝野對立", "藍綠對立", "藍綠惡鬥", "超越藍綠"]:
    tmp = merged.loc[:,["{}".format(kw),"year","month"]]
    tmp["ratio"] = tmp["{}".format(kw)].divide(merged["amount"])
    tmp = tmp.loc[:,["ratio","year","month"]].groupby(by=["year","month"], as_index=False).mean()
    tmp.rename(columns={"ratio":"ratio_{}".format(kw)}, inplace=True)
    ratio_month = pd.merge(ratio_month, tmp)

ratio_month = pd.merge(ratio_month, merged.loc[:,["year", "month", "amount"]], on=["year","month"])
ratio_month["ratio_sum"] = ratio_month.iloc[:, 2:7].sum(axis=1)
ratio_month.index = pd.date_range(start='01/01/1990', end='12/31/2018', freq='1M')
ratio_month = ratio_month.drop(["year", "month"],axis=1)
ratio_month
```

Out[6]:

	ratio_政黨惡鬥	ratio_朝野對立	ratio_藍綠對立	ratio_藍綠惡鬥	ratio_超越藍綠	amount	ratio_sum
1990-01-31	0.000000	0.000000	0.000000	0.000000	0.0	7960	0.000000
1990-02-28	0.000000	0.000231	0.000000	0.000000	0.0	8645	0.000231
1990-03-31	0.000000	0.000628	0.000000	0.000000	0.0	9549	0.000628
1990-04-30	0.000000	0.000000	0.000000	0.000000	0.0	9408	0.000000
1990-05-31	0.000000	0.000102	0.000000	0.000000	0.0	9794	0.000102
1990-06-30	0.000000	0.000301	0.000000	0.000000	0.0	9972	0.000301
1990-07-31	0.000000	0.000196	0.000000	0.000000	0.0	10217	0.000196
1990-08-31	0.000000	0.000000	0.000000	0.000000	0.0	9992	0.000000
1990-09-30	0.000000	0.000101	0.000000	0.000000	0.0	9906	0.000101
1990-10-31	0.000095	0.000474	0.000000	0.000000	0.0	10544	0.000569
1990-11-30	0.000000	0.000812	0.000000	0.000000	0.0	9854	0.000812
1990-12-31	0.000000	0.001063	0.000000	0.000000	0.0	9407	0.001063
1991-01-31	0.000000	0.000000	0.000000	0.000000	0.0	8885	0.000000
1991-02-28	0.000000	0.000000	0.000000	0.000000	0.0	7126	0.000000
1991-03-31	0.000000	0.000000	0.000000	0.000000	0.0	9367	0.000000
1991-04-30	0.000000	0.001727	0.000000	0.000000	0.0	8687	0.001727
1991-05-31	0.000000	0.001517	0.000000	0.000000	0.0	9231	0.001517
1991-06-30	0.000000	0.000115	0.000000	0.000000	0.0	8671	0.000115
1991-07-31	0.000000	0.000109	0.000000	0.000000	0.0	9158	0.000109
1991-08-31	0.000000	0.000415	0.000000	0.000000	0.0	9647	0.000415
1991-09-30	0.000000	0.001562	0.000000	0.000000	0.0	8324	0.001562
1991-10-31	0.000000	0.003466	0.000000	0.000000	0.0	9232	0.003466
1991-11-30	0.000000	0.001249	0.000000	0.000000	0.0	8810	0.001249
1991-12-31	0.000000	0.000435	0.000000	0.000000	0.0	9200	0.000435
1992-01-31	0.000000	0.000347	0.000000	0.000000	0.0	8639	0.000347
1992-02-29	0.000000	0.001472	0.000000	0.000000	0.0	6793	0.001472
1992-03-31	0.000116	0.000116	0.000000	0.000000	0.0	8608	0.000232
1992-04-30	0.000000	0.000334	0.000000	0.000000	0.0	8990	0.000334
1992-05-31	0.000000	0.000333	0.000000	0.000000	0.0	8998	0.000333
1992-06-30	0.000000	0.000000	0.000000	0.000000	0.0	8995	0.000000
...	...	...	...	...	...	...	...
2016-07-31	0.000371	0.000248	0.000619	0.000619	0.0	8077	0.001857
2016-08-31	0.000377	0.000000	0.000377	0.000126	0.0	7949	0.000881
2016-09-30	0.000134	0.000134	0.000134	0.000134	0.0	7480	0.000535
2016-10-31	0.000134	0.000403	0.000134	0.000134	0.0	7443	0.000806
2016-11-30	0.000553	0.000000	0.000553	0.000277	0.0	7228	0.001384
2016-12-31	0.000272	0.000408	0.000136	0.000544	0.0	7359	0.001359
2017-01-31	0.000000	0.000477	0.000636	0.000318	0.0	6288	0.001431
2017-02-28	0.000338	0.000000	0.000169	0.000000	0.0	5923	0.000507
2017-03-31	0.000000	0.000142	0.000000	0.000285	0.0	7019	0.000427
2017-04-30	0.000000	0.000160	0.000160	0.000000	0.0	6243	0.000320
2017-05-31	0.000587	0.000880	0.000440	0.001320	0.0	6820	0.003226
2017-06-30	0.000147	0.000442	0.000147	0.000147	0.0	6787	0.000884
2017-07-31	0.000147	0.000000	0.000000	0.001179	0.0	6786	0.001326
2017-08-31	0.000000	0.000000	0.000000	0.000145	0.0	6890	0.000145
2017-09-30	0.000146	0.000292	0.000146	0.000292	0.0	6855	0.000875
2017-10-31	0.000000	0.000000	0.000148	0.000148	0.0	6747	0.000296
2017-11-30	0.000296	0.000296	0.000000	0.000444	0.0	6753	0.001037
2017-12-31	0.000000	0.000000	0.000745	0.000149	0.0	6712	0.000894
2018-01-31	0.000305	0.000000	0.000305	0.000764	0.0	6548	0.001374
2018-02-28	0.000594	0.000000	0.000000	0.000198	0.0	5050	0.000792
2018-03-31	0.000428	0.000143	0.000000	0.000570	0.0	7015	0.001140
2018-04-30	0.000307	0.000000	0.000154	0.000307	0.0	6514	0.000768
2018-05-31	0.000730	0.000438	0.000730	0.000876	0.0	6852	0.002773
2018-06-30	0.001393	0.000000	0.000309	0.000928	0.0	6463	0.002630
2018-07-31	0.000308	0.000000	0.000154	0.000154	0.0	6490	0.000616
2018-08-31	0.000737	0.000147	0.000147	0.001180	0.0	6781	0.002212
2018-09-30	0.000157	0.000000	0.000157	0.000945	0.0	6351	0.001260

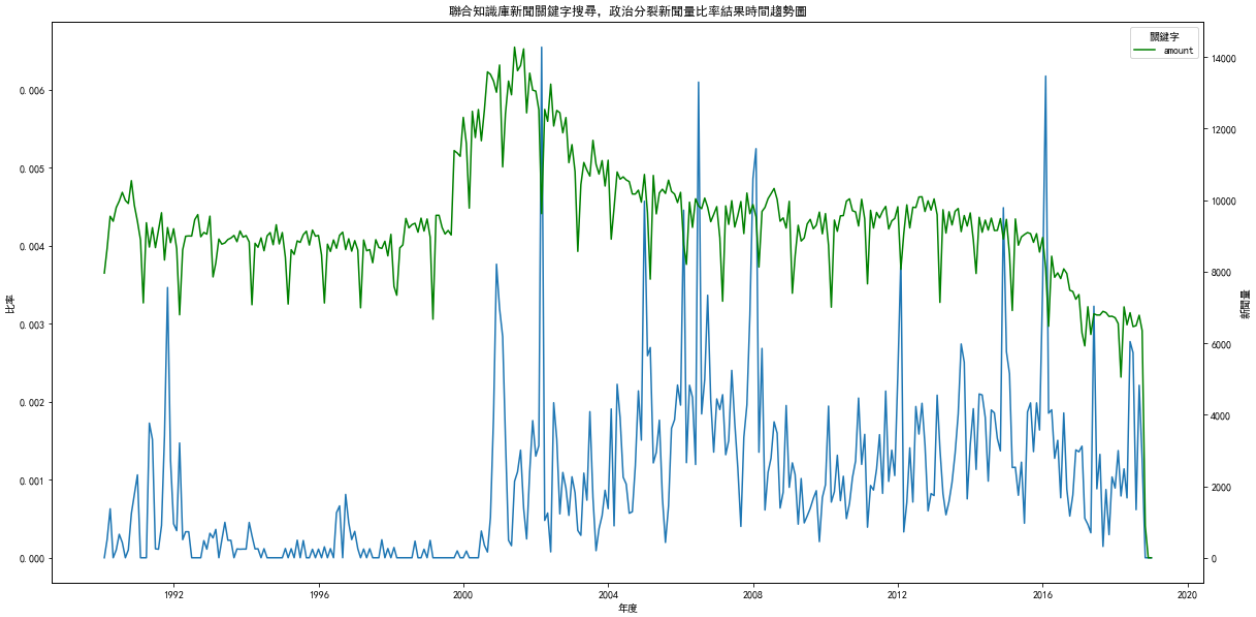
	ratio_政黨惡鬥	ratio_朝野對立	ratio_藍綠對立	ratio_藍綠惡鬥	ratio_超越藍綠	amount	ratio_sum
2018-10-31	0.000000	0.000000	0.000000	0.000000	0.0	890	0.000000
2018-11-30	NaN	NaN	NaN	NaN	NaN	0	0.000000
2018-12-31	NaN	NaN	NaN	NaN	NaN	0	0.000000

348 rows × 7 columns

畫圖

In [7]:
```
plt.rcParams["font.sans-serif"] = ["simhei"]
plt.figure(figsize=(20, 10))
plt.plot(ratio_month["ratio_sum"])
plt.ylabel("比率")
plt.xlabel("年度")
plt.twinx().plot(ratio_month["amount"],color="green")
plt.title("聯合知識庫新聞關鍵字搜尋，政治分裂新聞量比率結果時間趨勢圖")
plt.ylabel("新聞量")
plt.legend(prop={'size': 10}, title="關鍵字")
```

Out[7]:
```
<matplotlib.legend.Legend at 0x115cd4898>
```



In [11]:
```
ratio_month.plot(subplots=True, figsize=(20, 10),ylim=(0,0.005))
```

Out[11]:
```
array([<matplotlib.axes._subplots.AxesSubplot object at 0x116263748>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x1162a71d0>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x1162c3a20>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x1162deb38>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x1176fc0f0>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x117713668>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x11772bbe0>],
      dtype=object)
```