International Conference on on Soft Computing in Information Communication Technology (SCICT 2014)

Detecting Syphilis Amount in China Based on Baidu Query Data

Geng Peng
University of Chinese Academy of Sciences
Management school
Beijing, China
penggeng@ucas.ac.cn

Jiyuan Wang
University of Chinese Academy of Sciences
Management school
Beijing, China
Wangjiyuan12@mails.ucas.ac.cn

***Abstract*—Syphilis has drawn and is drawing more and more attentions globally because of its dangers and spreading speed, especially in China. Thanks to the development of search engine, a quicker and more accurate prediction of syphilis can be conducted. We collect the queries series on Baidu, a company providing search engine service in China. Several analyses are deployed to investigate the relationship between online search behaviors and the actual amount of the disease. Experiments show that accurate and fast predictions can be made using search queries. Finally, we also find that the recommendation of key words can increase the performance.**

Keywords-search data; search index; syphilis; prediction; key words recommondation

I. INTRODUCTION

Syphilis is a kind of sexually transmitted infectioncaused by the spirochete bacterum pallidum. In 1999, 12 million people were thought to be infected with syphilis, with greater than 90% of the amount in the developing world [1-3]. In china, CDC reports that tens of thousands of people are infected from 2007. Most importantly, this number does not include the infected who did not register in hospitals. The situation in China is more complicated since the immegration of rural peasants' crowd.

In the past, it's much harder to collect and track syphilitics because of people's concealing while today with the development of search engines; we can obtain the online search queries without disturbing those who are suffering from the disease. The idea of forcasting or nowcasting the offline activities using online behaviors is first introduced by Ginsberg et al [4]. Many related works are then produced afterwards, such as [5-9].

In this paper, we follow this idea, but care more about the queries selection strategy and its analysis. The google indices are not the real amounts of search frequency since they had been transformed by certain rules. But baidu queries we got are the actual numbers people search,so there is a demand for us to compound an index to make predictions, which is a big difference between our work and others on google.The second point is that the key words recommendation of Baidu will lead to similar key words sets, so there is a possible for us to investigate the problems by steps, rather than lacking of bound. Our work then can be repeated for commercial or academicals usage.So on theoretical and practical, our work has contributions in general. Especially for those who want to utilize the queries to conduct researches, a calculation saving strategy is a good choice.

In this paper, the size of query pool is considered as a new factor that may affect the performance of the prediction. The methodology we adopt is time series analysis. The idea implies that some part of the series can be predicted by the old information and the other can be predicted by the new variables considered then.The paper is organized as follows: Section II gives a review of literatures on the prediction of diseases like Syphilis and AIDS, prediction based on search queries, and some related methodologies also. In section III, we will discuss about the logic behind the online tracks. A general framework will be proposed in this section too. The results of empirical study will be given in Section IV. In section V, discussions are made about the effects of query pool size. Conclusions and future work can be seen in section VI.

II. LITERATURE REVIEW

Syphilis has drawn many attentions because of its harm to human beings; researches around this issue are generally divided into three categories. First one is the group who are investigating the underlying mechanics of the disease. The second group is trying to find some medicine, cure plans or the bad situations may have of syphilis. The last group of researchers is to study the behaviors of infected people, in order to make predictions or find spread patterns in the network [4-9].

In the first and second categories, many papers are published on this domain. Nicoll A, Hamers F F show that in Europe since 1995, the prevalence of gonorrhoea and syphilis, and that of HIV infection among heterosexuals, has been increasing [10].Franzen C reviewed biographies of several musicians and composers that probably suffered from syphilis [11]. Karp, G studies the co-infection of Syphilis and HIV [3]. Hall C S, Klausner J D, Bolan G A came up with a plan for those who were HIV-infected against syphilis [12]. Kent M E, Romanelli F reviewed the methodology about syphilis in 2008 [13].Coffin etc. cared about the problems that the syphilis problems in low-income developping countries [1]. Typically, Gao investigated the situations in China [2].

We focus on the third type. Several researches focused on the issues resent years. There exist many methods to make predictions: Autoregression analysis, using survey, using social network analysis and using search data etc. Among all the methods above, using search data seems to be more accurate.

Making search query based predictions are not new. Ginsberg et al first apply this method to make predictions on influenza epidemics [4]. Then sooner after that,D'Amuri etc. forecast the unemployment rate in US, and obtained a good result [5]. In the same year, a technique document was published by Google, aiming at predicting the future events with the search engine data [6]. N Askitas and KF Zimmermann also predicted the unemployment rate by using this idea [7]. In 2011, Zhi Da etc. applied this methodology to the finance domain also [8].Simeon Vosen and Torsten Schmidt forecast the private comsumption with two methods: survey and search index, eventually the latter turned out to be better [9].

Using search queries to make disease predictions is more and more popular in recent years. From the work of Ginsberg et al. to the current researches, new algorithms and models are proposed for many infectious diseases. All the jobs the researchers did are the fundamental attempts of this methodology. Especially in medical domain, a precise forecasting in a certain period of time is vital to both faculties and governments around the world. But during the previous researches, Chinese syphilis prediction is not conducted. Furthermore, some common sense in this domain has not been varified also. In this paper, we are trying to solve these two problems.

III. THEORETICAL FRAMEWORK

In this section, some basic ideas and logics of the relationship between the online and offline bahaviors are proposed for the preparation of future prediction and further discussion. The content latter are organized as followed. We first talk about the definition of search behavior, especially search for certain diseases. Then we draw a map for the patient wondering online and heading to hosipital for guidance. After the detail dicussion of the underlying casualty, a prediction model is raised for pratical usage and the notations in the model are described at the same time. Thirdly, a benchmark model is referred for comparison. Finally, some hypotheses and the key words selection strategies are proposed.

A. Conceptual Framework

The behavior of searching online is quite common in the internet era. It's a process people want to obtain exact information by organizing short texts, submitting the queries and assessing the results eventually. Generally, varied motivations drives people search through search engines. Most commonly, finding something a searcher does not know is its main purpose.

In the medical field, the usage of the search engine is much simple which mainly help the searcher to know related hospitals, diseases, symptoms etc. As to a typical pathema, the search queries are related to the corresponding symptoms or the name of the disease itself. Unfortunately, bias is exsited due to uncertainties we may not know.

It is necessary for us to think fully of the whole diagram concerning all the processes a patient may involves, ranging from finding symptoms to get cured finally. Fig 1 illustrated to details.

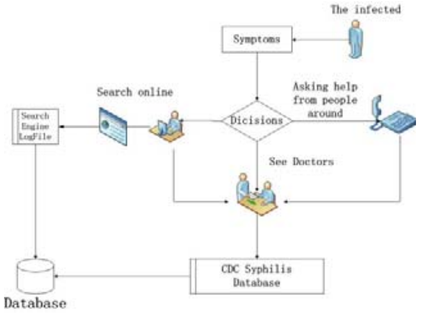


Figure 1. Diagram of decision-making process form a syphilis patient

When someone suspecting itself a syphilis-like patient, the first thing it will do is to confirm whether their symptoms are like the syphilis ones. Till now, several paths to make clear of the symptoms are most adopted by those disease-like people. The first choice is to get help from doctors in hospitals, while commonly before seeing doctors, people prefer to ask some details about the symptoms in some informal ways. So asking help from friends or neighbors and searching online are the most used methods. Because syphilis is generally related to personal affairs, searching online seems to be a secure and fast way. After the simple check, the infected will go to hospitals for professional treatment. Then the registration will be written to the log file of hospitals.

From the description above, an expalanatory conceptual logic between the query numbers and the actual number has been established. Even that, we cannot prove that the universal rule really exists, so statistical experiments are necessary to test the relationship in a probabilistic perspective.

B. Model for prediction

Basically, simple time series analysis is enough for our prediction use right now. The main reason is that the characteristics of the series we consider are quite stationary, well-structured and very close to linear ones. The details of the series will be demonstrated in the next section.

Like most diseases, syphilis amount in each month is volatilly increasing and with certain seasons. So the prediction model should have following factors considered.

The first one is the search queries which this paper proposes to use. The second is the lagged dependent vaiable, thus the amount of the disease. Finally, a controlled variable is considered is the holiday, because in China, Spring Festival is the national holiday that all the people are involved, so in that if a month contains the holiday, we must make a tag. So the general model is like equation (1):

$$S_t = \beta_0 + \beta_1 \cdot S_{t-1} + \beta_2 \cdot Index + \delta \cdot holiday \qquad (1)$$

In the model, represents the actual amount of syphilis in each month, and stands for the lagged series. Term Inc is the series we composite standing for the whole queries. Term $holi_t$ marks the special month. Finally, is an error term.

Considered the index, the method we propose is a heuristic hierarchy method. Details of the key words selection is demonstrated in a tree shown in Fig.2.

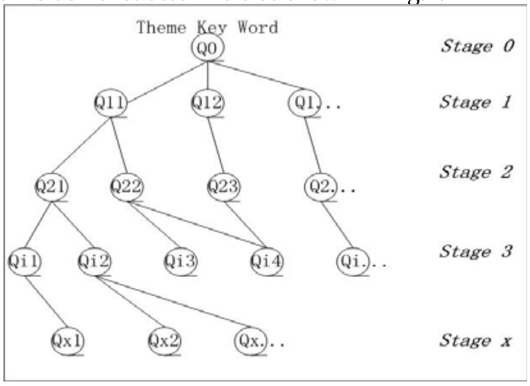


Figure 2. Key words selection process

In the first selection stage, we choose the name of the disease, in this paper, which is “syphilis”. In stage 2, we use the recommendations of the query “syphilis”, and make a query pool. In the latter satges, same strategies are applied. Besides, duplicate elimination is conducted. The reason is that the query in the former stages may also be recommended too. We ensure that the same query appear in the query pool just one time.

C. *Benchmark model*

It is not convincing without comparisons.To avoid that situation, benchmark models should be referred also to make comparisons.

The benchmark models are without the index term or the holiday term, which are as follows in equation (2) to (4):

$$S_t = \beta_0 + \beta_1 \cdot S_{t-1} + \varepsilon \quad . \ (2)$$

$$S_t = \beta_0 + \beta_1 \cdot S_{t-1} + \delta \cdot \quad (3)$$

$$S_t = \beta_0 + \beta_1 \cdot S_{t-1} + \beta_2 \quad (4)$$

To assess the performance of the prediction, an accuracy measure should be considered. The MAPE formula in shown in equation (5):

$$MAPE = \frac{1}{h+1} \quad (5)$$

D. *Theoretical hypothesis*

To make our prediction more convincing, we need to provide several extended experiment outcomes in the following sections. In this part of the section, we just raised some hypotheses.

Firstly, we assume that with the increasing of the pool size, the performance should be better than those with smaller pool size. The basic idea is that we suppose new related informantion can decrease the probability that the predicted value has a larger distance to actual value. The corresponding relationship is focused on the index and the size of the pool.

Secondly, the holiday may have a negative effect on the amount of syphilis each month. This is because we just mentioned that the faculty of the registration department may have days off. Even though this situation may not exist, during that day, patients may also delay the examination time concerning the great event.

Finally, we think the effect of increasing the pool size is limited. This may be seen on the decreasing of the changing rate of the improvement of the performance.

IV. EMPIRICAL STUDY

In this section, several experiments will be conducted to inllutrate the performances of the model we proposed in previous sections.Firstly, we will introduce the data source and related preprocess. Secondly, key words selection strategies and compositing approach will be given according to the tree in section V. Then, to prove the stationeriness, stationary tests will be manipulated. Finally, the results of the models will be demonstrated in a comparatively way.

A. *Data Source*

All the data sets are from baidu and CDC, respectively. The queries are the day size, and the CDC data is on months, so we roll up the query data of baidu. Furthermore, we mark out the spring festival, and add a sequence called holiday, the reason we did so is that in Spring Festival, the registration staff were all taking days off.

Baidu Index’s database (http://index.baidu.com/) contains logs of online search queryvolume submitted from June 2006. However, since the influenza case count data isavailable from March 2009, we use Baidu’s data from March 2009 to August 2012.

Besides, the range of the data is from January 2007 to March 2013. The last 4 months’ data are choosen to be a testing set for assessing the predicting ability and the others are used for estimating the parameters of the models.

Fig.2 shows the curve of the syphilis data.

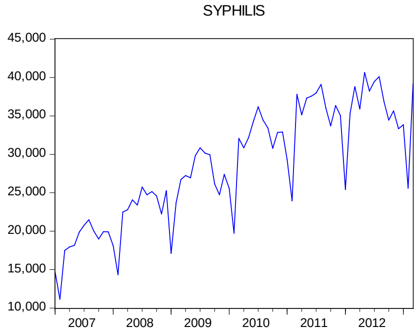


Figure 3. Curve of syphilis amount in China

B. *Keywords Selection and Index Compositing*

Different Queries may have different search amount and can therefore produce diversemodeling results. Keywords are carefully Chosen to reflect terms most likelyassociated with syphilis.We use keywords recommended by search engine and the process is decribed as following.

In this paper, we extend our key word selecting stage to 5th level. That is to say, we compound 5 indexes gradually in each stage. The theme key word in stage 0 is “syphilis”. Because of length limit, key words in first 2 stages are demonstrated in Tab.1.

Tab. 1 shows the keywords after simple filtering.

TABLE I. SERACH QUERY SELECTION IN EACH STAGE

Query pool in Stage 0	梅毒	
Query pool in Stage 1	梅毒	淋病
	梅毒症状	梅毒的早期症状
	梅毒能治愈吗	性病
	美国梅毒恐慌	梅毒治疗
	梅毒的治疗	

The key point of the approach above is that the approach is repeatable and with less subjective. For the methods of many researchers on this area, subjective collecting is sometimes biased and time-consuming. The idea of using recommendation of search engine itself totally can be automated. So a faster prediction can be made to help dicision makers.

Therefor, some key words in Tab.1 seem unreasonable but useful. However, we can imagine the situations in a patient’s perspective. For example, gonorrhoeais is also referred in Tab.1. The reason is that once someone got secual diseases, he or she may not know what kind of sexual disease he or she has got. So searching for familiar disease names is possible, and then he or she will contrast the sympotons with his or hers.

Is worth noting, Chinese is quite different from English. In Chinese, a same meaning may have many different explanations and expressions. So it is hard for us to find a fix algorithm to fit all the situations by applying simple natuaral language processing methods. Expert and recommendation are the two useful approaches because they do not have to deal with the key words or the content in the queries.

The next few steps focus on the how the queries composite an index representing the overall characteristics of the queries. In stage 0, the only one key word stands for the index itself. In the other stages, a compounding approach is needed.

We adopt an approach called lagged compounding approach to make these processes realized. The main idea of this method is using greedy strategy to fulfill the maixmum the closeness beyween syphilis series and the index series. The steps of the method are as follows:

Firstly, we calculate the correlation between the queries with different lags and syphilis series. The maximum lag should not exceed a seasonal cycle according to expert experiences. For example, according to the graph of the series, a rough lag l is obtained. Then $l$correlations will be calculated for each query.

Secondly, the lagged series with largest correlation with syphilis series will be picked up for every query in the query pool.

Thirdly, summing up all the laggede series to an index series is the final step of the index compounding in a certain stage.

Finally, we repeat all the steps to compound an index in a new stage. Iterations do not end untill all the stage are finished.

Based on this approach, we compound five indexes from stage 0 to stage 4. The curves are shown in Fig.4.

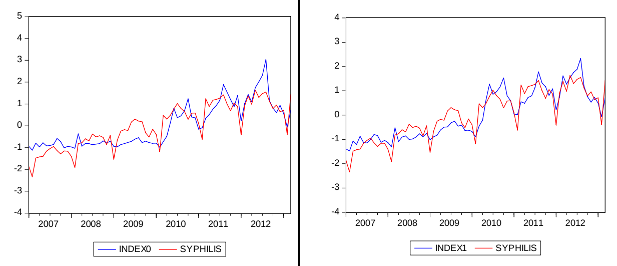
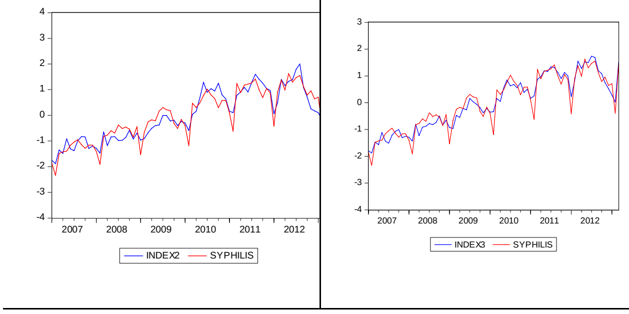


Figure 4.a Syphlis and index0

Figure 4.b Syphlis and index1

Figure 4.c Syphlis and index2

Figure 4.d Syphlis and index3



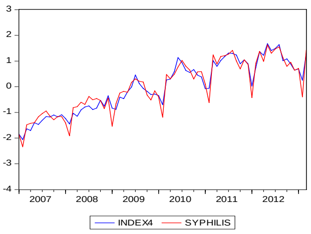


Figure 4.d Syphlis and index4

Figure 4. Normalized curves of syphilis amount and indexes in stages

It can be seen that, with icreasing of the size of query pool, the fitness of the index and syphilis series grows. In the next experiments, we will verify this proposition.

C. Co-intergrationTest

Unit root tests are conducted for stationary tests in this paper. The extended Dickey - Fuller test method (referred to as the ADF test) are adapted, the null hypothesis is that there exists at least one unit root for the series, and it means the series is not stable. The stationary test results are shown in Tab.2 and 3. It can be concluded that the original series of three variables are not stable, but the first difference of them are stable at the 1% significance for rejecting the original hypothesis, indicating that they are all first order stationay series.

TABLE II. RESULTS OF SYPHILIS' STATIONARY TEST

Note: △ represents first difference

			t-Statistic	Prob.*
Syphilis Amount	**Augmented Dickey-Fuller test**		-2.672	0.0846
	Test critical values:	**1% level**	-3.538	
		5% level	-2.908	
		10% level	-2.592	
△Syphilis Amount			**t-Statistic**	**Prob.***
	Augmented Dickey-Fuller test		-7.089	0
	Test critical values:	**1% level**	-3.538	
		5% level	-2.908	
		10% level	-2.592	

TABLE III. RESULTS OF INDEXES' STATIONARY TEST :P-VALUE

Note: △ represents first difference

All the p-value is at 1% level

	Raw data	**△**
Index0	**0.3162**	**0.0001**
Index1	**0.3409**	**0.0000**
Index2	**0.3059**	**0.0000**
Index3	**0.3472**	**0.0000**
Index4	**0.0992**	**0.0000**

In Table 2 and 3, all the series of variables are not stationary at 1% level, which means that no strong evidence we have to reject the original hypothesis. But with first order differences, all the series are stationary with ADF tests.

Based on the co-integration theory, using time series data which are the non-stationary originally but with the same co-integration order, the results are stable in estimation and prediction.

D. *Prediction*

Five test models are employed to make predictions, the benchmark models are set to be compared. We first use our model to examine the performances of five different models with different indexes in corresponding stages. The results can be seen in Table 4. The controllable variables are the lagged syphilis term and the holiday binary-state variable.

In Table 4, all the variables pass the F-tests, which means they are significantly non-zeros. And the equations satisfy all the statistical conditions like R^2, AIC, SC and so forth at the highest level. The results can be seen in Table 4 and Figure 5.

Is worth noting, we estimate our models using samples from January 2007 to November 2012. And the testing sets are from December 2012 to March 2013. The MAPE measures in the table are of the testing sets. Model (1) to (5) represents the model with index0 to index4, respectively.

TABLE IV. RESULTS COMPARISON OF MODELS BETWEEN INDEXES

Month	**Actual Amount**	**Model (1)**	**Model (2)**	**Model (3)**	**Model (4)**	**Model (5)**
2012-12	33332	37014.79	36265.22	34078.67	33750.1	34315.86
2013-1	33851	30680.77	30870.75	29857.39	30866.85	33033.1
2013-2	25594	29804.24	28874.46	27549.38	25099.01	27220.62
2013-3	39190	38942.45	38148.33	38900.76	39493.37	38501.16
MAPE		9.37%	8.27%	5.60%	3.19%	3.28%
R square		0.924	0.935	0.946	0.961	0.967

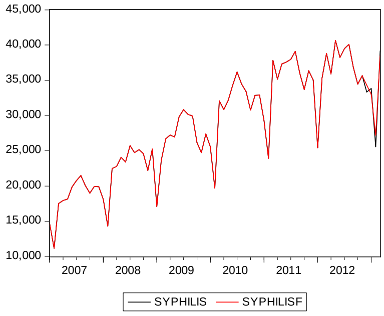


Figure 5. Actual and predicted value

We need further to verify that the effects of variable of holiday. Then the results of models without holiday and index will be demonstrated in Table 5 and Table 6.

TABLE V. RESULTS COMPARISON OF MODELS II

Month	**Actual Amount**	**Model without hopliday or index**	**Model without index**
2012-12	33332	37462.69	37413.23
2013-1	33851	29270.28	29835.99
2013-2	25594	37687.2	31881.14
2013-3	39190	40677.52	40386.65
MAPE		19.24%	12.93%
R square		0.852	0.910

A simple glance of the results help us understand the improvement of the holiday event variable and the co-intergration outcome. Further discussion and managerical implications will be referred in details in next section.

V. DISCUSSION

According to the results in previous section, some interesting implications can be found.

Firstly, we verified that the adding of new information can help improve the predicting performance, which can be seen in Table 4. Generally, the new relevent new information can increase the probability we win in guessing that the events will happen or not.

Secondly, the size of query pool can help improve the performance also. As we can see in Table 4, with the increasing of the pool size, the performance appears to be better except model 5 with index4. To clarify that this outcome should be blamed for a mistake or a rule, we then list the $R^2$of each model. No decreasement can be seen in fitness of model 5.

Thirdly, the holiday variable is reasonable and useful. At the beginning, we propose to use this variable by our common sense, which is lack of scientific proof. Then we consider it as a controllable variable, and the results show that it's significant in statistics. In case that it's an occasional situation, we set a benchmark model to examine it. So from all the description and experiments above, holiday should be considered in the model.

Finally, the usage of lagged term of syphilis series is reasonable. The lag of the term is 12, thus syphilis (-12) is considered in our model, which turns out to be significant also. From the graph we also can see that the seasoned cycle is nearly 12 months.

VI. CONCLUSIONS

In this paper, a forcasting model is built to predict the syphilis amounts. In the early section, previous works conducted by researchers around the globe are reviewed and the lacking of study in the high performance prediction is referred. In the next few parts, we discussed the paradigm of the online (especially the search behaviors)-offline relationship, meanwhile a framework is proposed to illustrated the logic behind the two behaviors. In section III, time series analysis forcasting model was raised for prediction. At the same time, a benchmark model which is an AR model was referred for a contrastive study. In the empirical study, we first collected the data on Baidu and CDC. Then the results are given and the comparison shows that our model has a higher accuracy. In the discussion section, we further studied the effects of adding new information from internet which turned out to be that the new information can help increase the prediction performance. Furthermore, a study of the effect of query pool size on prediction performance is conducted, and the results show that firstly, with the increasing of the size query pool, the performance is gradually improved. Secondly, the changing rate of the performance is reducing while the size is growing, thus the improvement can convergence to a point where we treat the pool correspondingly reaches maximum.

Although the model turned out to be a better model, some problems emerge also. Firstly, the recommendation of queries is based on the algorithms the search engine service provider uses. So a further discussion on this topic requires more attention. Finally, the proposed model is an offline model due to its complex preprocesses and processes. So a faster learning and forcasting algorithm is expected to help the decision adoptors acquire the new information faster.

ACKNOWLEDGMENT

This research is supported by the National Natural ScienceFoundation of China under Grant 71202115, 71172199 and71203218,and we would like to thank their support.

REFERENCES

[1] Coffin, LS; Newberry, A, Hagan, H, Cleland, CM, Des Jarlais, DC, Perlman, DC (January 2010). "Syphilis in Drug Users in Low and Middle Income Countries". The International journal on drug policy 21 (1): 20–7.

[2] Gao, L; Zhang, L, Jin, Q (September 2009). "Meta-analysis: prevalence of HIV infection and syphilis among MSM in China". Sexually transmitted infections 85 (5): 354–8.

[3] Karp, G; Schlaeffer, F, Jotkowitz, A, Riesenberg, K (January 2009). "Syphilis and HIV co-infection". European journal of internal medicine 20 (1): 9–13.

[4] Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detectinginfluenza epidemics using search engine Query data", Nature,2008, 457(7232),pp. 1012–1014.

[5] D'Amuri, F. and J. Marcucci, "'Google it!' Forecasting the US unemployment rate with a Google jobsearch index", MPRA paper, University Library of Munich, Germany, 2009.

[6] Choi, H. and H. Varian. (2009). Predicting the present with Google trends. Technical report, 2009b, Google Inc.

[7] N Askitas and KF Zimmermann. (2009). Google Econometrics and Unemployment Forecasting. Applied Economics Quarterly, 2009, 55(2), 107-120.

[8] Zhi Da, Joseph Engelberg, Pengjie Gao,(2011). In Search of Attention, The Journal of Finance, 2011,66(5): 1461–1499.

[9] Simeon Vosen,Torsten Schmidt,(2011). Forecasting private consumption: survey-based indicators vs. Google trends. Journal of Forecasting, 30(6): 565–578.

[10] Nicoll A, Hamers F F. Are trends in HIV, gonorrhoea, and syphilis worsening in western Europe?[J]. BMJ: British Medical Journal, 2002, 324(7349): 1324.

[11] Franzen C. Syphilis in composers and musicians—Mozart, Beethoven, Paganini, Schubert, Schumann, Smetana[J]. European journal of clinical microbiology & infectious diseases, 2008, 27(12): 1151-1157.

[12] Hall C S, Klausner J D, Bolan G A. Managing syphilis in the HIV-infected patient[J]. Current infectious disease reports, 2004, 6(1): 72-82.

[13] Kent M E, Romanelli F. Reexamining syphilis: an update on epidemiology, clinical manifestations, and management[J]. The Annals of pharmacotherapy, 2008, 42(2): 226-236.