Mining of Web-Page Visiting Patterns with Continuous-Time Markov Models

Qiming Huang[1], Qiang Yang[2], Joshua Zhexue Huang[3], and Michael K. Ng[4]

[1]College of Computer Science & Technology, Beijing University of Post & Telecom
qm_huang@yahoo.com
[2]Department of Computer Science, Hong Kong University of Science & Technology
[3]E-Business Technology Institute, The University of Hong Kong
[4]Department of Mathematics, The University of Hong Kong

Abstract. This paper presents a new prediction model for predicting when an online customer leaves a current page and which next Web page the customer will visit. The model can forecast the total number of visits of a given Web page by all incoming users at the same time. The prediction technique can be used as a component for many Web based applications . The prediction model regards a Web browsing session as a continuous-time Markov process where the transition probability matrix can be computed from Web log data using the Kolmogorov's backward equations. The model is tested against real Web-log data where the scalability and accuracy of our method are analyzed.

Keywords: Web mining, Continuous Time Markov Chain, Kolmogorov's backward equations, Sessions, Transition probability

1 Introduction

Web mining is a thriving technology in the practice of Web-based applications. By applying data mining technologies such as clustering, association rules and discrete Markov models, Web mining has been successfully applied to Web personalization and Web-page pre-fetching. Study has shown that the next page an online customer is going to visit can be predicted with statistical models built from Web sessions [4, 5]. However, an open problem is *when* an online customer will click on a predicted next page. A related question is *how many* customers will click the same page at the same time. In this paper, we answer these questions from a perspective of Markov models.

Markov models are one of major technologies for studying the behaviors of Web users. In the past, discrete Markov models are widely used to model sequential processes, and have achieved many practical successes in areas such as Web-log mining. The transition matrix based on the Markov process can be computed from visiting user-session traces, and the Frobenius norm of the differences of two transition probability matrices can show the difference of the two corresponding sequences [4]. Users can be clustered by learning a mixture of the first-order Markov models with an Expectation-Maximization algorithm [7]. The relational Markov model (RMMs) makes effective learning possible in domains of a very large and heterogeneous state space with only sparse data [11].

H. Dai, R. Srikant, and C. Zhang (Eds.): PAKDD 2004, LNAI 3056, pp. 549–558, 2004.

In this paper, we present a continuous-time Markov model for the prediction task. In contrast with the previous work which uses the prediction rules [5], we use the Kolmogorov's backward equations to compute the transition probability from one Web page *A* to another Web page *B* according to the following steps: first, we preprocess the Web-log data to build user sessions; second, we compute the transition rate of a user leaving a Web page *A* and obtain a transition rate matrix; third, we compute the transition probability matrix using the Kolmogorov's backward equations. From the transition probability matrix we can predict Web page a user will visit next, and when the user will visit the page. Furthermore, we can find the total transition count from all other Web pages to the predicted Web page at the same time.

The main contribution of this work is to put forward two hypotheses for computing a transition probability model from one Web page to another. The first hypothesis treats Web browsing sessions as a continuous time Markov process. The second hypothesis regards the probability of leaving a Web page as having an exponential distribution over the time. We can compute the transition rate for leaving the current Web page with the second hypothesis, and then according to the first hypothesiswe compute the transition probability matrix by the Kolmogorov's backward equations.

The paper is organized as follows. Section 2 presents the prediction model. Section 3 presents the experiment results. Section 4 concludes the paper.

2 Continuous Time Markov Chains for Prediction Models

A Web log often contains millions of records, where each record refers to a visit by a user to a certain Web page served by a Web server. A session is a set of ordered Web pages visited in one visit by the same visitor at a given time period.

A sequence of pages in a user session can be modeled by a Markov chain with a finite number of states [5, 9]. In discrete Markov chains, we need to consider the minimal time interval between page transitions, which is not easy to predict. In this paper, we propose to use continuous-time Markov chains for predicting the next visiting web page and when, and the total transition count from all other Web pages to the predicted Web page, instead of using discrete Markov chains. As a result, we can manage different time intervals in which to visit Web pages.

2.1 Continuous Time Markov Chains

A stochastic process is called a *continuou`time Markov chain at* state *i*

- The amount of time it spends in state *i*, before making a transition to a different state, is *exponentially distributed* with rate v_i, and
- It enters the next state *j* from state *i* with probability P_{ij}, where $P_{ii} = 0$ and $_j P_{ij} =$ 1 for every possible state *j* in the state space.

Let { $X(t), t \geq 0$ } be a continuous-time Markov chain of a user browsing Web pages. Here the state of a continuous-time Markov chain refers to a Web page and the state space contains all possible Web pages, which is finite but large. The main characteristics of a continuous-time Markov chain is that the conditional distribution of the next Web page to be visited by a user at the time *t+s* is only dependent on the present Web page at the time *s* and is independent of the previous Web pages. For simplicity, we let

$$P_{ij} = P[\, X(t+s) = j \mid X(s) = i \,].$$

In a small time interval *h*,

- $P_{ii}(h)$ is the probability that the process in state *i* at time 0 will not be in state *i* at time *h*. $P_{ii}(h) = v_i\, h + o(h)$ is equal to the probability that a transition occurs in (*0, h*).
- $P_{ij}(h) = hv_i P_{ij} + o(h)$ is the probability that a transition occurs in (*0, h*) from state *i* to state *j*.

The limits of $(1 - P_{ij}(h)) / h$ and $P_{ij}(h) / h$ are equal to v_i *and* $v_i P_{ij}$ respectively as *h* approaches zero.

In order to predict the time at which a user will visit the next Web page, we make use of the well-known equation *Kolmogorov's Backward Equations* [1] in the continuous-time Markov chain:

$$P_{ij}'(t) = v_i \times \sum_{k \neq i} P_{ik}(t) \times P_{kj}(t) - v_i \times P_{ij}(t)$$

The transition rate matrix $R\{\, r_{ij} \}$ with its entries is defined as

$$r_{ij} = v_i \; x \; P_{ij} \;\; if \; i \neq j$$
$$r_{ij} = -v_i \;\; if \; i=j.$$

The Kolmogorov backward equations can be rewritten as

$$P'_{ij}(t) = \Sigma \; r_{ik} P_{kj}(t).$$

We can write it in the matrix form as

$$P'(t) = RP(t).$$

This is a system of ordinary differential equations and the general solution is given by

$$P(t) = e^{Rt}.$$

The above formula gives the probability of the next Web pages to be visited at time *t*. We need to compute the matrix *R* in order to use the continuous-time Markov chain for the prediction.

2.2 Computation of the Transition Rate Matrix *R*

According to the definition of P_{ij} *and* v_i in the previous subsection, the term v_i is the rate at which the Markov process leaves Web page *i* and P_{ij} is the probability that the Markov process enters Web page *j*.

Let us use the popular NASA Web-log data set to demonstrate how to compute v_i *and* P_{ij}. To illustrate, we extract a URL which index is 2327. Firstly, we compute the time t_{ij} ($j=1,2,\ldots,n$) for all people who visited the Web page *i* and entered another page *j* after the visit, where *n* is the number of visiting next Web pages from Web page i. Then we sort all t_{ij} (t_{i1}<t_{i2}<...t_{il}). Next we compute the accumulative frequency N_{im} of visits of Web page i up to time t_{im}. We note that N_i is the accumulative frequency of visits of Web page i at the time the last visitor leaves. Therefore, we can calculate the probability of leaving aWeb page i as N_{im}/N_i at time t_{im}.

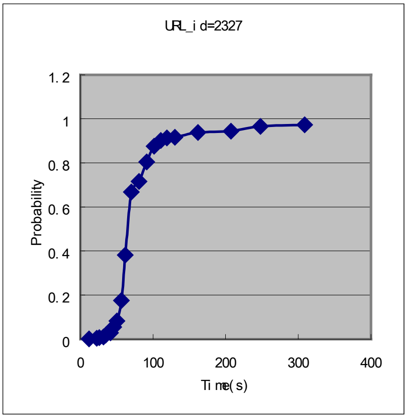

Fig. 1. The probability distribution of Leaving a Web page with time

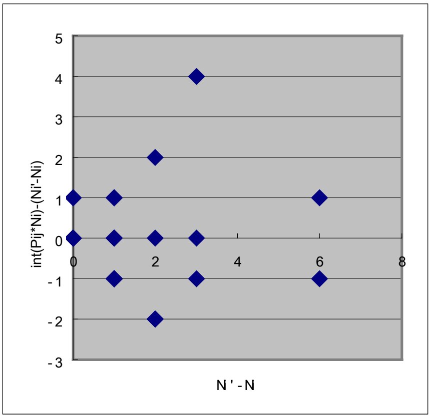

Fig. 2. The Deviation of the Actual Count Visiting Page 2334 from the predicted count

Figure 1 shows that the probability leaving from a Web page is exponentially distributed. By considering the statistical hypothesis in the continuous-time Markov chain, we model the curve in Figure 1 by an exponential distribution function as follows:

$$F(t)=1-e^{-\lambda t},\ \ t\ \ 0\ \ and\ \ F(t)=0,\ \ t<0$$

Where λ is equivalent to the transition rate v_i of the continuous-time Markov chain [1]. λ can be determined as follows:

$$\lambda\ = -(\ ln\ (1-F(t)\ )\ /\ t.$$

Based on this formula, we estimate the transition rate v_i as follows:

$$v_{im} = -\ (ln\ (\ 1-N_{im}/\ N_i)\ )\ /\ t_{im} \qquad m=1,2,...,n-1$$
$$v_i = (v_{i1} + v_{i2}+\ ...+\ v_{i(n-1)})\ /\ (n-1).$$

For each Web page, we employ the same procedure to obtain the estimates of the transition rate v_i. Finally, the probability P_{ij} of leaving Web page i to Web page j can be estimated from the data by counting the relative frequency of the next visit to Web page j from Web page i. Finally, we obtain the transition rate matrix R from P_{ij} and v_i by the follow formula.

$$r_{ij}= v_i\ x\ P_{ij}\ \ if\ \ i \neq j$$
$$r_{ij}= -v_i\ \ if\ \ i=j.$$

3 Empirical Analyses

3.1 Experimental Setup

The experiment was conducted on the NASA data set, which contains one month worth of all HTTP requests to the NASA Kennedy Space Center WWW server in Florida. The log was collected from 00:00:00 August 1, 1995 through 23:59:59 August 31, 1995. We filtered out documents that were not requested in this experiment. These were image requests or CSS requests in the log that were retrieved automatically after accessing requests to a document page containing links to these files and some half-baked requests [5].

We consider the Web log data as a sequence of distinct Web pages, where subsequences, such as user sessions can be observed by unusually long gaps between consecutive requests. To save memory space, we use number IDs to identify Web pages and users. To simplify the comparing operation, the time has been transformed to seconds starting from 1970.

In deciding on the boundary of the sessions, we make up two rules as follows.
In a user session, if the time interval between two consecutive visiting is larger than 1800 seconds, we consider the next visit starting a new session. If a user has two consecutive records visiting the same Web page in 1800 seconds, we consider the next visit to be a new session. We loaded all sessions into a session data cube and operated the cube to extract the set of sessions for building the continuous-time Markov chain as described in Section 3.

3.2 Computing the Transition Probability from One Web Page to Others

After the continuous-time Markov chain for visiting Web pages was built, we used formula $P(t) = e^{Rt}$ to calculate the probability of entering another Web page from the current page. The Matrix $P(t)$ is estimated by:

$$P(t) = e^{Rt} = \lim_{n \to \infty} (I + Rt/n)^n$$

To obtain the limiting effect, we raised the power of the matrix $I+Rt/n$ to the nth power for sufficiently large n.

3.3 Predicting Which Next Pages People Will Visit, and When

We ran experiments to test the validity of the prediction model. Tracing through the NASA data set, a person visiting the Web page 2359 will decide where to go next. We make a prediction on when and where this visitor will go next, using our model.

Let us set the start time when the person was visiting the Web page 2359 as zero second. We count the transition count from the page 2359 to other pages before the start time of prediction. A total of 19 pages were visited by the person who left the page 2359 before the zeroth second; these pages are listed in Table1, where N_{ij} is the transition count from Web page $i=2359$ to Web page j that actually happened in the data set before the zeroth second.

Table 1. The transition parameter from the Web page 2359 to other Web pages in 600 seconds.

URL_id$_j$	2324	2442	2383	2364	2334	2362	2969	2333	2358	2496	2375	2344	2325	2372	2657	2374	2393	2370	2327
N$_{ij}$	1	1	2	1	32	1	5	7	1	2	2	1	3	1	2	1	1	1	
p$_{ij}$	0.0093	0.0013	0.0493	0.0111	0.0367	0.0343	0.0049	0.0193	0.0650	0.0714	0.0237	0.0130	0.0234	0.0321	0.0018	0.0168	0.0011	0.0026	0.0259
N'$_{ij}$	1	1	2	1	33	1	5	7	1	2	2	1	3	1	2	1	1	1	

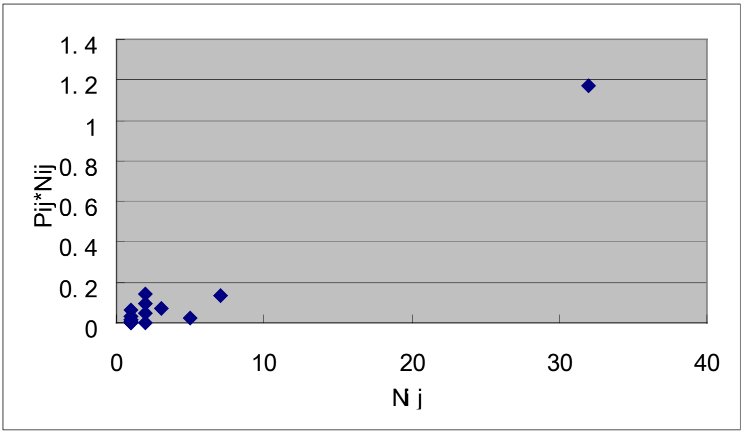

Fig. 3. The $P_{ij} \times N_{ij}$ value at 600 second

With our model, we can compute the transition probability values P_{ij} from Web page 2359 to all other Web pages at time *t*. When the time *t* is 600 seconds, Figure 3 shows the value of $P_{ij} \times N_{ij}$, where the integer part is the predicted transition count from Web page i to Web page j during a given time interval. The value of int($P_{ij} \times N_{ij}$) at Web page 2334 increases up to one, and we can predict that the next page to go from Web page 2359 is Web page 2334 and the transition time is within 600 seconds.

We use N'_{ij} to denote the total transition count from Web page *2359* to another Web page *j* that actually happened in the NASA data set before the end of prediction time. In Table 1, N'_{ij} - N_{ij} is one at Web page 2334. The person left Web page 2359 to Web page 2334 in 600 seconds in reality.

3.4 Predicting the Visiting Count of a Web Page

With the prediction time increasing from zero second and up, we can predict the next Web page visited by an online person, when int($P_{ij} \times N_{ij}$) value first becomes one. We note the corresponding time visiting the next Web page as t_1 seconds. N_i is the total transition count from Web page *i* to other Web pages before the start of prediction. Int($P_{ij} \times N_i$) is the predicted count of visitors leaving Web page *i* and entering Web page *j* during the prediction time. Thus, we can compute the total transition count from all Web pages to the designation Web page in the interval of zero to t_1 seconds.

Table 2. The transition parameters from all Web pages to the Web page 2334 in 600 secs.

URL_id$_i$	2324	2539	2352	2327	2442	2356	2383	2388	2440	2387
P$_{ij}$	0.036	0.045	0.035	0.038	0.011	0.038	0.037	0.043	0.024	0.047
N$_i$	214	4	74	187	25	9	108	39	2	19
int(P$_{ij}$*N$_i$)	7	0	2	7	0	0	3	1	0	0
N$_i$′	220	4	75	190	26	9	111	40	2	21
N$_i$′-N$_i$	6	0	1	3	1	0	3	1	0	2

URL_id$_i$	2362	2333	2358	2496	2375	2322	2379	2349	2441	2438
P$_{ij}$	0.037	0.033	0.029	0.037	0.04	0.044	0.041	0.03	0.039	0.043
N$_i$	60	38	26	15	101	46	38	67	14	13
int(P$_{ij}$*N$_i$)	2	1	0	0	4	2	1	2	0	0
N$_i$′	62	38	27	15	103	47	39	70	15	13
N$_i$′-N$_i$	2	0	1	0	2	1	1	3	1	0

URL_id$_i$	2669	2325	2372	2359	2357	2374	2473	2329	2423	3131
P$_{ij}$	0.028	0.035	0.021	0.037	0.047	0.046	0.031	0.042	0.035	0.044
N$_i$	6	150	22	65	9	38	16	25	25	7
int(P$_{ij}$*N$_i$)	0	5	0	2	0	1	0	1	0	0
N$_i$′	6	156	22	66	9	38	16	25	26	7
N$_i$′-N$_i$	0	6	0	1	0	0	0	0	1	0

URL_id$_i$	2990	2338	2400	2389	2506	2517	2505	2376	2419	2682
P$_{ij}$	0.035	0.015	0.033	0.052	0.042	0.033	0.035	0.040	0.050	0.046
N$_i$	7	10	28	3	3	7	14	24	14	12
int(P$_{ij}$*N$_i$)	0	0	0	0	0	0	0	1	0	0
N$_i$′	7	10	29	3	3	8	16	25	14	13
N$_i$′-N$_i$	0	0	1	0	0	1	2	1	0	1

URL_id$_i$	2336	2405	2439	2854	2447	2600	2953	2411	2462	2504
P$_{ij}$	0.049	0.044	0.020	0.023	0.018	0.038	0.022	0.049	0.029	0.026
N$_i$	26	20	6	5	9	3	5	4	9	4
int(P$_{ij}$*N$_i$)	1	0	0	0	0	0	0	0	0	0
N$_i$′	26	20	6	5	9	3	5	4	9	4
N$_i$′-N$_i$	0	0	0	0	0	0	0	0	0	0

URL_id$_i$	2561	2918	2558	2578	2468	2672	2767	2637	2699	3129
P$_{ij}$	0.033	0.081	0.024	0.032	0.029	0.061	0.029	0.026	0.022	0.040
N$_i$	2	2	1	4	16	2	5	3	4	2
int(P$_{ij}$*N$_i$)	0	0	0	0	0	0	0	0	0	0
N$_i$′	2	2	3	5	16	2	5	3	4	2
N$_i$′-N$_i$	0	0	2	1	0	0	0	0	0	0

Table 2 shows the result of all previous 60 pages from which visitors entered the page 2334 within 600 seconds. The second row lists the transition probability values (P$_{ij}$) from another page to page 2334 in 600 seconds. N$_i$ (or N$_i$$'$) is the total visiting count from the previous Web page (URL_id$_i$) to Web page 2334 before the start (or the end) of the prediction time. N'_j - N_i is the actual visiting count from the Web page (URL_id$_i$) to Web page 2334 during the prediction time. The total count of visiting Web page 2334 in 600 seconds was calculated as 45 using Σabs(P$_{ij}$ x N$_i$) and the actual visiting count was 42 according to Σabs(N$_i$$'$ −N$_i$).

Figure 3 shows the prediction deviation value (int(P_{ij} x N_i)-(N$_i$$'$ −N$_i$)), which mostly vary in the scope of [-1,1]. The error rate of the total predicted visiting count to a Web page is computed as follows.

$$ER_j = \left(\sum_{i=1}^{m} ER_{ij} \right) / m$$

where

$$ER_{ij} = \begin{cases} 0 & if\ \mathrm{int}(P_{ij} \times N_i) = (N_i' - N_i) \\ 1 & if\ \mathrm{int}(P_{ij} \times N_i) \neq 0\ and\ (N_i' - N_i) = 0 \\ \dfrac{abs(\mathrm{int}(P_{ij} * N_i) - (N_i' - N_i))}{N_i' - N_i} & if\ \mathrm{int}(P_{ij} \times N_i) \neq (N_i' - N_i)\ and\ \mathrm{int}(P_{ij} \times N_i) \neq 0 \\ & and\ 0 < abs(\mathrm{int}(P_{ij} \times N_i) - (N_i' - N_i)) < \mathrm{int}(P_{ij} \times N_i) \\ 1 & if\ \mathrm{int}(P_{ij} \times N_i) \neq (N_i' - N_i)\ and\ \mathrm{int}(P_{ij} \times N_i) \neq 0 \\ & and\ abs(\mathrm{int}(P_{ij} \times N_i) - (N_i' - N_i)) \geq (N_i' - N_i) \end{cases}$$

The error rate of the predicted count in visiting Web page 2334 in Table 2 was calculated as 34.4%. The main reason for the error rate is that some people do not visit the Web page with the pattern in the training data set. In the prediction model, the computation of the transition probability P_{ij} as shown in the following formula may be unstable when the parameter n is not large enough. This causes an error in the prediction.

$$P(t) = e^{Rt} = \lim_{n \to \infty} (I + Rt/n)^n$$

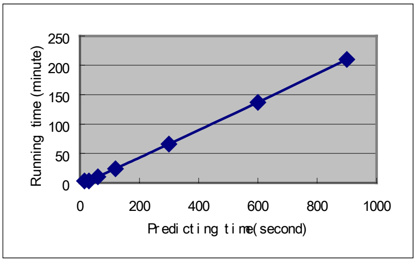

Fig. 4. The relation of the run time and the predicting time.

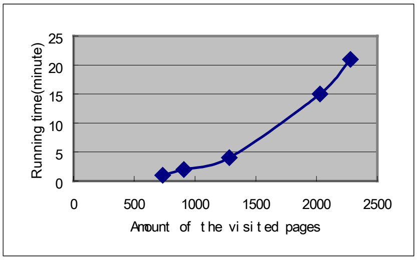

Fig. 5. The run time to the amount of the visited pages

3.5 Scalability Experimental Results

We conducted some experiments to test the scalability of the continuous Markov chain prediction method. The experiments were carried out on a Pentium III, 662 MHz with 196MB RAM. The NASA data in 39608 seconds from 00:00:00 August 1, 1995 was used. We computed the transition probability in different predicting time length (*T*). Figure 4 shows the linear relation of the predicting time length and the computing time. For a certain precision value, the relation of the predicting time length (*T*) and the running time is linear too.

Using a different size of training data sets which included different numbers of visited Web pages, we computed the transition probability within the next ten seconds. Figure 5 shows the relation of the number of visited Web pages and the running time.

4 Conclusions

In this paper, we explored using the continuous-time Markov chain to predict not only the next Web page a person will visit and the time when it will happen, but also the visiting count of the Web page in the same time. The transition probability, computed from the continuous-time Markov chain, gives us rich information from which we can compute the transition count from one Web page to another, as well as the total number of visits to a Web page by all people within a certain period of time. The prediction model is validated by an experiment. In the future, we plan to continue to explore the application of continuous time Markov models in Web page prediction. We will also consider how to apply the prediction result to prefetching applications.

Acknowledgement. Michael Ng, Joshua X. Huang, and Qiang Yang are supported by grants from HK ITC and RGC.

References

[1] William J. Anderson (1991) Continuous Time Markov Chains: An Applications-Oriented Approach. Springer-Verlag, New York.

[2] C.M.Leung and J.A.Schormans (2002) Measurement-based end to end latency performance prediction for SLA verification. 41st European Telecommunications Congress, Genoa

[3] J.M.Pitts and J.A.Schormans (2000) Introduction to IP and ATM design and performance. 2nd edition, Wiley.

[4] Qiang Yang, Joshua Zhexue Huang and Michael Ng (2002) A Data Cube Model for Prediction-based Web Prefetching. Journal of Intelligent Information Systems, Vol. 20 (2003), 11-30.

[5] Qiang Yang, Hui Wang and Wei Zhang (2002) Web-log Mining for Quantitative Temporal-Event Prediction. IEEE Computer Society, Computational Intelligence Bulletin, Vol. 1, No. 1, December 2002.

[6] James Pitkow and Peter Pirolli (1999) Mining Longest Repeating Subsequences to Predict World Wide Web Surfing. In Proceedings of USENIX Symposium on Internet Technologies and Systems. 1999.

[7] Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, Steven White (2000) Visualization of Navigation Patterns on a Web Site Using Model-Based Clustering. In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, USA.

[8] Mukund Deshpande and George Karypis (2001) Selective Markov Models for Predicting Web-Page Accesses. In Proceedings SIAM Int. Conference on Data Mining.

[9] Taha, H. (1991) Operations Research, 3rd Edition, Collier Macmillan, N.Y., USA.
[10] Sheldon M. Ross (1996) Stochastic Process. Wiley.
[11] Corin R. Anderson, Pedro Domingos, Daniel S. Weld (2002) Relational Markov Models and Their Application to Adaptive Web Navigation. In Proceedings of SIGKDD 2002.
[12] Alan Jennings and J.J. Mckeown (1992) Matrix Computation. John Wiley & Sons.