IDENTIFICATION OF THE DYNAMICS OF THE GOOGLE'S RANKING ALGORITHM

A. Khaki Sedigh, Mehdi Roudaki

Control Division, Department of Electrical Engineering,
K.N.Toosi University of Technology
P. O. Box: 16315-1355, Tehran, Iran
sedigh@eetd.kntu.ac.ir, roudaki@iranseo.com

Abstract: Among the search engines, Google is one of the most powerful. It uses an accurate ranking algorithm to order web pages in search results. In this paper, it is shown that a simple linear model can approximately model the dynamics governing the behaviour of Google. Least Squares is used for the system identification procedure. Identification results are provided to show the effectiveness of the identified system.

Keywords: Google Search Engine, Ranking Algorithm, System Identification, Least Squares.

1. INTRODUCTION

Search engines are tools to help web users in finding their favourite information or web sites. Today most people start at a search engine on the web. Popularity of search engines depends on their accurate search results. The more accurate the search results are, the more popular the search engine is. Search engines have become increasingly important on the web and their ranking procedure is a fundamental characteristic which is important to web sites.

E-commerce is much attuned to the ranking issue, because higher ranking translates directly into more sales. A top ranking in a major search engine often will generate more targeted traffic than an expensive banner advertising campaign. Plus, a good search engine position is free-anyone can do it (Marchini, 1999). Today, Google is one of the most important search engines. It is one of the most important places to be well listed. The resulting traffic can be staggering (Jimworld, 2002).

Google uses an accurate ranking algorithm to order web pages in its search results. Its ranking algorithm is unknown and is not available to public.

In order to compare the web pages and order them according to their relevance to the searched term, Google considers many parameters for web pages which are also unknown to the Google users.

In this paper, the dynamics of the ranking algorithm of Google is considered as a black box. The inputs of which are the parameters collected from the web pages according to the searched term and the outputs are the ranks of those web pages which the parameters are calculated for them in Google search results. Simulation results are provided in the paper to show the identified model capabilities to predict the Google's behaviour in its ranking algorithm.

2. LEAST SQUARES SYSTEM IDENTIFICATION

Least Squares is a well-known and established technique for the identification of the dynamical behaviour of linear dynamical plans. It is assumed that the computed variable $\hat{y}$ is given as (Astrom and Wittenmark, 1995):

$$\hat{y} = \theta_1\varphi_1(x) + \theta_2\varphi_2(x) + ... + \theta_n\varphi_n(x) \quad (1)$$

Where $\varphi_1, \varphi_2, ..., \varphi_n$ are known parameters and

$\theta_1, \theta_2, ..., \theta_n$ are the unknown parameters. Observation pairs $\{(x_i, y_i), i = 1, 2, ..., N\}$ are obtained from the experiments performed and the parameters are determined to minimize the following cost function:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{N}\varepsilon_i^2 \quad (2)$$

where

$$\varepsilon_i = y_i - \hat{y}_i \quad (3)$$

In a compact form, the solution to the least squares method problem is given by

$$\boldsymbol{\theta} = (\boldsymbol{\varphi}^T\boldsymbol{\varphi})^{-1}\boldsymbol{\varphi}^T\mathbf{y} \quad (4)$$

where

$$\boldsymbol{\varphi} = \begin{bmatrix} \boldsymbol{\varphi}^T(x_1) \\ : \\ \boldsymbol{\varphi}^T(x_N) \end{bmatrix} \quad (5)$$

is assumed fullrank and

$$\boldsymbol{\varphi}^T(x_1) = [\varphi_1\ \varphi_2 \ldots \varphi_n]^T \quad (6)$$

also,

$$\mathbf{y} = [y_1\ y_2 \ldots y_N]^T \quad (7)$$

3. IDENTIFICATION FRAMEWORK

Google's search results are not fixed and the web pages ranks always change. Changes in search results are caused by factors such as: indexing new web pages by Google, effort of webmasters to optimize their web sites for higher rankings, the dynamic property of Google's algorithm and probably few bugs in Google ranking algorithm.

In this paper, terms are searched in different time cycles then the stored search results are compared in order to select web pages which their ranks mostly vary between 1 and 100. To guarantee this condition, 87 is the worst considered rank. This kind of data collecting is for considering the steady state of ranks not the transient ranks of some web pages.

The next step is to select the number of observations. By observing sum square error given by equation (2) for many different cases, the number of the web pages was finally chosen as 26.

In selecting the web pages for the identification process, the following points are considered:

- All search results web pages are ignored if their URLs (uniform resource locator) have changed.
- The softwares can calculate the web pages parameters.
- Most of the parameters of some pages are zero.
- None of the parameters of some web pages are not zero.
- Some of sequential web pages have the same PageRank (this seems to be essential because it helps to clarify other parameters influences. PageRank is the Google's criteria in measuring the importance of web pages).
- Number of data sets (N) is bigger than unknown parameters (n).

3.1 Selection of the Input Parameters

For identification purposes, an important step is the selection of the number of input parameters. Many resources state that Google uses more than 100 parameters. (Wen, 2002) points that the main factors in Google search engine ranking algorithm are:

- Anchor text (the text of a link or hyperlink) from Yahoo and Dmoz
- Anchor text
- PageRank
- Keyword proximity (the placement of words on a web page in relation to each other)

And finally states that anchor text is the most important factor in Google's ranking algorithm.

Generally there are many reports about Google's parameters but so far there is no such confirmed data regarding the Google's parameters. Therefore in this paper based on the experiences of the authors these input parameters are considered:

- PageRank
- Keyword frequency (how often a word appears in a web page) in the web pages visible text.
- Keyword density (the number of words appearing on a web page compared to the total number of words appearing on that web page) in the web pages title.
- Keyword density in the web pages text.
- Keyword density in the web pages linked text.

- Keyword density in the web pages ALT tags (the ALT tag is a label describing an image.).
- Keyword prominence (how early in a web page's text a word appears) in the web pages text.

And these places are considered in Similar Pages (when a user select the Similar Pages link for a particular result, Google automatically scouts the web for pages that are related to this result):

- The title of Similar Pages.
- Text below the title (this text is an excerpt from the returned result page with the query terms bolded.) of Similar Pages.
- The category of Similar Pages in ODP (Open Directory Project).
- The URL of Similar Pages.

To calculate the parameters which are related to the Similar Pages a binary form is considered:

- If there is the searched term in it, it is considered as 1.
- If there isn't the searched term in it, it is considered as 0.

This kind of valuation method is shown in table 1.

Table 1 A binary form is considered to calculate the parameters which are related to the Similar Pages. It is assumed that the searched term is "search engine optimization". "Q" means the searched term is available exactly. Q means at least one word of searched term is available. NC is abbreviation of Not Considered.

Case	"Q"	Q
URL: www.searchenginewatch.com	NC	1
URL: www.pandia.com/searchworld/	NC	1
URL: www.toika.com/optimization/	NC	1
URL: www.rankwrite.com	NC	0
THE TITLE AND THE TEXT BELOW THE TITLE: search engine optimization	1	1
THE TITLE AND THE TEXT BELOW THE TITLE: search engine placement and optimization	0	1
THE TITLE AND THE TEXT BELOW THE TITLE: optimization	0	1
CATEGORY: Computers>Internet>…>Search Engine Optimization Firms>S	NC	1
CATEGORY: Computers>Internet>…>Promotion>News	NC	0
CATEGORY: Computers>Internet>…>Free Search Engine Submitting	NC	1

After collecting and normalizing the input data the equation (4) is solved and the ranks are calculated by equation (1).

Fig.1 shows the observed ranks, the calculated ranks are shown in Fig.2, the observed and calculated ranks are compared in Fig.3 and Fig.4 shows the errors in the estimation of ranks.

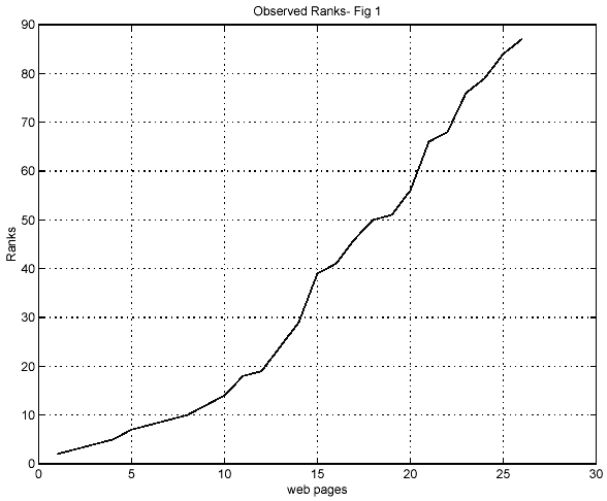

Fig. 1. Observed ranks.

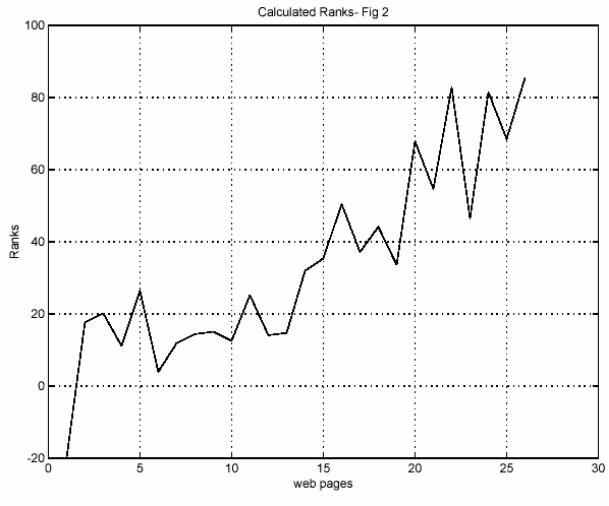

Fig. 2. Calculated ranks.

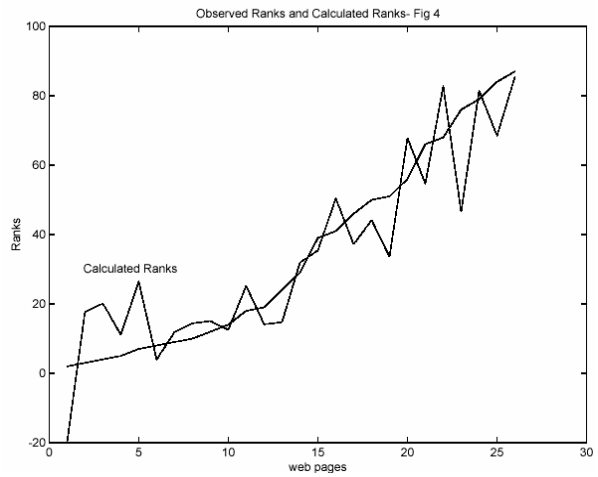

Fig. 3. Observed and calculated ranks are compared.

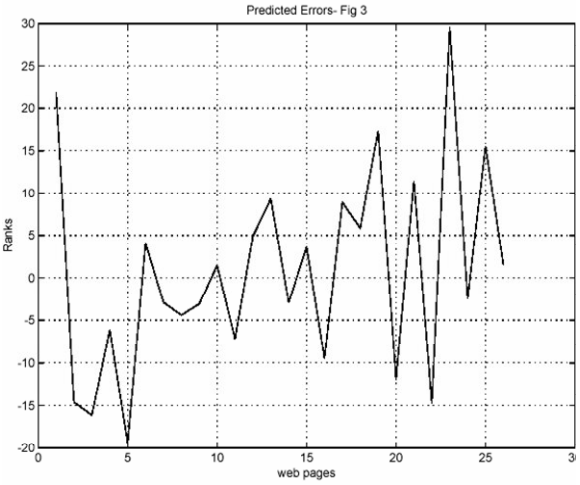

Fig. 4. Errors in the estimation of the ranks.

3.2 Analysis of the Identified System

To test the accuracy of the identified system, the parameters of the 5 web pages are calculated and then their ranks are estimated. These 5 web pages are not used in the identification calculations. Table 2 shows the result of this analysis.

Table 2 Analysis of the accuracy of the identified system

Ranks in Google's search results	Calculated ranks	Difference
1	-34	35
35	51	16
37	32	5
70	65	5
82	53	29
90	65	25

The ranks of the web pages which are used in the system identification are as follows:

$$2 \leq \text{rank} \leq 87 \qquad (8)$$

As it is shown in table 2, the identified system can predict the ranks which are between 2 and 87 with an acceptable error but can not predict the ranks which are out of this region.

4. CONCLUSION

Least squares identification is employed to identify the dynamics behaviour governing the ranking algorithm of Google search engine. It is shown that an accurate identification of Google's ranking algorithm is not possible, because Google does not show the entire web pages which are related to the searched term and its parameters are not known. However an approximate linear model is found to describe the ranking dynamics.

It is shown that Google's dynamics can be partially modelled using a linear model. Identification results using real data are presented to show the main points of the paper.

REFERENCES

Astrom, K.J. and B. Wittenmark (1995). *Adaptive Control*, 41-89, Addison Wesley, New York.

Jimworld. (2002). Search engine forums at Jimworld, http://www.searchengineforums.com.

Marchini, F. (1999). Secret's to achieving a top 10 position, First place software Inc.

Wen, P. (2002). Google search engine ranking algorithm analysis, Pwqsoft Inc., http://www.pwqsoft.com/search-engine-ranking.htm.