Markov Chain Monte Carlo Ⅱ

Woojoo Lee

Consider the information retrieval problem:

Given a query string

1) Compute the similarity scores of each documents

2) Rank them with the scores

Now, we consider another information retrieval problem for webpages.

However, it would be naïve to apply the information retrieval method for documents to webpages if we do not consider a big difference between them.

→ Links between webpages !

How can we assign the PageRank score to each webpage ?

The basic idea of PageRank algorithm:

Weight the links from different webpages according to

1) High weight on high PageRank score

2) Less weight if the webpage has many links to other webpages

Some notations are introduced:

Consider n webpages.

1. $L_{ij} = 1$ if webpage j links to webpage i, otherwise 0
2. $m_j = \sum_{k=1}^{n} L_{kj}$

According the previous two principles, we can derive

The ith PageRank score=

We can understand the PageRank problem clearly when we write the above formula with matrix notations.

Interpreting PageRank as a Markov Chain

1) The webpages → states

2) Find the transition probability matrix

3) What is the meaning of the PageRank score under Markov Chain framework ?

When the PageRank score vector is well-defined ?
(existence and uniqueness)

Consider the following transition probability matrix :

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Q) Is the PageRank score vector is well-defined ?

→ How can we resolve this problem ?

A modified solution is given by

$$\mathrm{P}(\text{go from } i \text{ to } j) = \begin{cases} (1-d)/n + d/m_i & \text{if } i \rightarrow j \\ (1-d)/n & \text{otherwise} \end{cases}$$

Find the transition probability matrix.

How can we find the PageRank for very large n ?

1. An efficient iterative method

2. Sparsity of the transition probability matrix