

Метод расчета рейтинга популярности веб-страниц

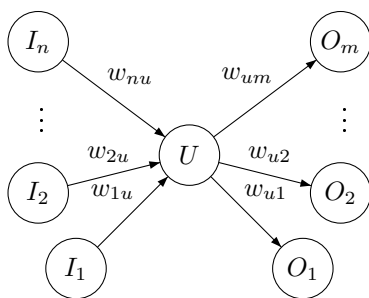
Максим Захаров <dr.maxime@gmail.com>

29 Января 2007*

В этой статье описывается метод расчета рейтинга популярности *Neo*, используемый в поисковом движке *DataparkSearch*¹.

При использовании метода *Neo* подразумевается, что страницы являются нейронами, а ссылки между страницами – связями между нейронами. Таким образом, для образованной нейронной сети можно использовать один из методов обучения – метод обратного распространения ошибки. Рейтингом популярности страницы в данном случае считается уровень активности соответствующего нейрона.

Для простоты вычислений, каждый проиндексированный документ (*U*) рассматривается отдельно, и вместе с документами, ссылающимися на него (I_1, I_2, \dots, I_n), а также документами, на которые он ссылается (O_1, O_2, \dots, O_m), образуют однослойную нейронную сеть:



Здесь $w_{1u}, w_{2u}, \dots, w_{nu}$ – веса входящих ссылок и $w_{u1}, w_{u2}, \dots, w_{um}$ – веса исходящих ссылок.

В качестве активационной функции стандартно используется сигмоидальная функция

$$f(x) = \frac{1}{1 + e^{-x}}$$

*Last-Modified: 11 Апреля 2007

¹См. <http://www.dataparksearch.org/>

Реальным выходом нейрона *U* является

$$o_u = f\left(\sum_{k=1}^n w_{ku} \cdot Neo(I_k)\right),$$

где $Neo(X)$ – значение рейтинга популярности для нейрона *X*.

Ожидаемым выходом нейрона *U* будем считать

$$d_u = f\left(\sum_{k=1}^m w_{uk} \cdot Neo(O_k)\right).$$

Если реальный и ожидаемый выход нейрона не совпадают, то производится коррекция весов входящих связей следующим образом:

$$\Delta w_{uk} = -c \cdot (d_u - o_u) \cdot o_u \cdot (1 - o_u) \cdot Neo(O_k), k = 1 \dots n,$$

где *c* – постоянный коэффициент скорости обучения [1, Глава 10].

В общем случае, число итераций для схождения $o_u \rightarrow d_u$ не ограничено какой-либо величиной, поэтому после конечного числа итераций, задаваемого в качестве параметра алгоритма, значение *Neo*-рейтинга для нейрона *U* полагается равным

$$Neo(U) = \frac{d_u + O_u}{2}.$$

Описанные выше шаги расчета рейтинга популярности одного нейрона последовательно применяются для всех проиндексированных документов коллекции.

Если документ не имеет входящих или исходящих ссылок, в этом случае реальный или ожидаемый выход нейрона полагается равным некоей малой величине ϵ .

Список литературы

- [1] Люгер, Джордж, Ф. Искусственный интеллект: стратегии и методы решения сложных проблем, 4-е издание, Пер. с англ. М. Издательский дом "Вильямс 2003. ISBN 5-8459-0437-4