

Способы учета свойств вершин графа при подсчете PageRank

Максим Жуковский

MSU, MIPT, Yandex

18 сентября 2013

- Мотивация

- Мотивация
- Алгоритм PageRank

- Мотивация
- Алгоритм PageRank
- Способы учета свойств вершин

- Мотивация
- Алгоритм PageRank
- Способы учета свойств вершин
 - Алгоритм Semi-Supervised PageRank

- Мотивация
- Алгоритм PageRank
- Способы учета свойств вершин
 - Алгоритм Semi-Supervised PageRank
 - Алгоритм BrowseRank

- Мотивация
- Алгоритм PageRank
- Способы учета свойств вершин
 - Алгоритм Semi-Supervised PageRank
 - Алгоритм BrowseRank
 - Алгоритм Semi-Supervised BrowseRank

- Мотивация
- Алгоритм PageRank
- Способы учета свойств вершин
 - Алгоритм Semi-Supervised PageRank
 - Алгоритм BrowseRank
 - Алгоритм Semi-Supervised BrowseRank
 - Алгоритм Nested Semi-Supervised PageRank

- Мотивация
- Алгоритм PageRank
- Способы учета свойств вершин
 - Алгоритм Semi-Supervised PageRank
 - Алгоритм BrowseRank
 - Алгоритм Semi-Supervised BrowseRank
 - Алгоритм Nested Semi-Supervised PageRank
- Применения и эксперименты

- Ранжирование страниц. Данные: выборка страниц, размеченных по запросам.

- Ранжирование страниц. Данные: выборка страниц, размеченных по запросам.

Факторы:

- Ранжирование страниц. Данные: выборка страниц, размеченных по запросам.

Факторы:

- Основанные на графовой структуре веба.

- Ранжирование страниц. Данные: выборка страниц, размеченных по запросам.

Факторы:

- Основанные на графовой структуре веба.
 - PageRank

- Ранжирование страниц. Данные: выборка страниц, размеченных по запросам.

Факторы:

- Основанные на графовой структуре веба.
 - PageRank
- Запросо-зависимые.

- Ранжирование страниц. Данные: выборка страниц, размеченных по запросам.

Факторы:

- Основанные на графовой структуре веба.
 - PageRank
- Запросо-зависимые.
 - Запросо-зависимый PageRank

- Ранжирование страниц. Данные: выборка страниц, размеченных по запросам.

Факторы:

- Основанные на графовой структуре веба.
 - PageRank
 - Запросо-зависимые.
 - Запросо-зависимый PageRank
-
- Нахождение спама. Данные: про некоторые страницы точно известно, что они являются спамом, про другие — что они не являются спамом.

- Ранжирование страниц. Данные: выборка страниц, размеченных по запросам.

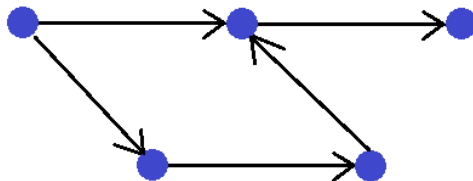
Факторы:

- Основанные на графовой структуре веба.
 - PageRank
 - Запросо-зависимые.
 - Запросо-зависимый PageRank
-
- Нахождение спама. Данные: про некоторые страницы точно известно, что они являются спамом, про другие — что они не являются спамом.

PageRank — один из факторов.

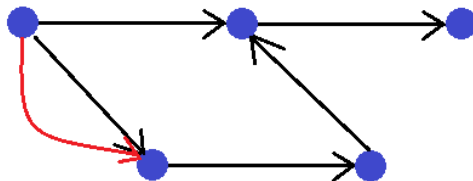
Пусть N — количество страниц в интернете.

$$\text{PR}(p) = \frac{0.15}{N} + 0.85 \sum_{\tilde{p} \rightarrow p} \frac{1}{\#\{p' : \tilde{p} \rightarrow p'\}} \text{PR}(\tilde{p}).$$



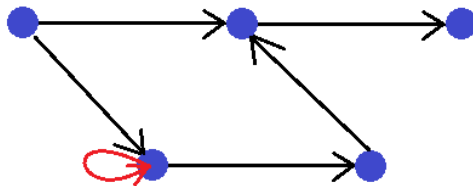
Пусть N — количество страниц в интернете.

$$\text{PR}(p) = \frac{0.15}{N} + 0.85 \sum_{\tilde{p} \rightarrow p} \frac{1}{\#\{p' : \tilde{p} \rightarrow p'\}} \text{PR}(\tilde{p}).$$



Пусть N — количество страниц в интернете.

$$\text{PR}(p) = \frac{0.15}{N} + 0.85 \sum_{\tilde{p} \rightarrow p} \frac{1}{\#\{p' : \tilde{p} \rightarrow p'\}} \text{PR}(\tilde{p}).$$



$$\pi_{i+1} = A\pi_i,$$

$$A = (a_{i,j}), \quad a_{i,j} = p(i \rightarrow j),$$

$$p(i \rightarrow j) = \frac{\alpha}{N} + \frac{1 - \alpha}{\text{outdegree}(i)},$$

$$\alpha = 0.15.$$

- V — множество вершин, E — множество ребер;
- $\mathbf{V} = (\mathbf{V}_i, i \in V)$ — факторы вершин, $\mathbf{E} = (\mathbf{E}_{ij}, i, j \in V)$ — факторы ребер,
 $\mathbf{V}_i$ — l -мерный вектор, содержащий факторы вершины i ,
 $\mathbf{E}_{ij}$ — h -мерный вектор, содержащий факторы ребра $i \rightarrow j$;
- $f(\omega, i)$ — вес вершины i , $g(\phi, i \rightarrow j)$ — вес ребра $i \rightarrow j$.

Semi-Supervised PageRank [B. Gao, T.-Y. Liu, W. Wei, T. Wang, H. Li, *Semi-Supervised Ranking on Very Large Graphs with Rich Metadata*, KDD'11]:

$$\pi(i) = \alpha \frac{f(\omega, i)}{\sum_{h \in V} f(\omega, h)} + (1 - \alpha) \sum_{\tilde{i}: \tilde{i} \rightarrow i \in E} \frac{g(\varphi, \tilde{i} \rightarrow i)}{\sum_{j: \tilde{i} \rightarrow j \in E} g(\varphi, \tilde{i} \rightarrow j)} \pi(\tilde{i}).$$

$$f(\omega, i) = \langle \omega, \mathbf{V}_i \rangle, \quad g(\varphi, i \rightarrow j) = \langle \varphi, \mathbf{E}_{ij} \rangle.$$

$$\pi_{i+1} = A\pi_i,$$

$$A = (a_{i,j}), \quad a_{i,j} = p(i \rightarrow j),$$

$$p(i \rightarrow j) = \alpha \frac{f(\omega, j)}{\sum_{h \in V} f(\omega, h)} + (1 - \alpha) \frac{g(\varphi, i \rightarrow j)}{\sum_{\tilde{j}: i \rightarrow \tilde{j} \in E} g(\varphi, i \rightarrow \tilde{j})}.$$

- $f_q(p)$ — значение фактора вершины p для запроса q .
- Обучающее множество: множества вершин $V_q^1, V_q^2, \dots, V_q^k$ (упорядоченные от наиболее релевантных до наименее релевантных вершин).
- $h(i, j, f_q(p_2) - f_q(p_1))$, $p_1 \in V_q^i, p_2 \in V_q^j$, — значение штрафа, где $b_{ij} > 0$, $1 \leq i < j \leq k$,
 $h(i, j, x) = \max\{x + b_{ij}, 0\}^2$.

Оптимизационная задача:

$$F(\omega) = \sum_q \sum_{1 \leq i < j \leq k} \sum_{p_1 \in V_q^i, p_2 \in V_q^j} h(i, j, f_q(p_2) - f_q(p_1)) \rightarrow \min.$$

$$\pi_{i+1} = A\pi_i,$$

$$\pi_{i+1} = A\pi_i,$$

$$\left(\frac{\partial \pi}{\partial \chi}\right)_{i+1} = \frac{\partial A}{\partial \chi} \pi_i + A \left(\frac{\pi}{\chi}\right)_i.$$

- Сессии.

- Сессии.
- Продолжительность пребывания на странице.

- Сессии.
- Продолжительность пребывания на странице.
- Вес страницы, вес ребра.

- Сессии.
- Продолжительность пребывания на странице.
- Вес страницы, вес ребра.
- Среднее время пребывания на странице вычисляется исходя из того, что время пребывания равно сумме шума, имеющего распределение χ -квадрат, и экспоненциального распределения.

- Сессии.
- Продолжительность пребывания на странице.
- Вес страницы, вес ребра.
- Среднее время пребывания на странице вычисляется исходя из того, что время пребывания равно сумме шума, имеющего распределение χ -квадрат, и экспоненциального распределения.

π_i — стационарное распределение в непрерывном случае,

$\tilde{\pi}_i$ — стационарное распределение в дискретном случае,

$p(i \rightarrow j)$ — вероятности переходов в дискретной цепи,

$Q(i)$ — среднее время пребывания на странице.

$$p(i \rightarrow j) = (1 - \alpha)\sigma_j + \alpha \frac{w_{ij}}{\sum_{k=1}^{N+1} w_{ik}}, i \in V, \quad p(N+1 \rightarrow j) = \sigma_j,$$

$$\pi_i = \tilde{\pi}_i Q(i) / \left(\sum_{j=1}^N \pi_j Q(j) \right).$$

$$p^{\text{SSP}}(i \rightarrow j) = (1 - \alpha)\sigma_j^{\text{SSP}} + \alpha \frac{w_{ij}^{\text{SSP}}}{\sum_{k=1}^{N+1} w_{ik}^{\text{SSP}}}, i \in V,$$

$$p^{\text{SSP}}(N+1 \rightarrow j) = \sigma_j^{\text{SSP}},$$

$$\sigma_j^{\text{SSP}} = \sigma_j f(\omega_i), \quad w_{ij}^{\text{SSP}} = w_{ij} g(\varphi, i \rightarrow j),$$

$$f(\omega, i) = \langle \omega, \mathbf{V}_i \rangle, \quad g(\varphi, i \rightarrow j) = \langle \varphi, \mathbf{E}_{ij} \rangle.$$

$$\omega = (\omega_1, \dots, \omega_l, \tilde{\omega}_1, \dots, \tilde{\omega}_l), \quad \varphi = (\varphi_1, \dots, \varphi_h, \tilde{\varphi}_1, \dots, \tilde{\varphi}_h).$$

$$\begin{aligned} \omega &= (\omega_1, \dots, \omega_l), \quad \tilde{\omega} = (\tilde{\omega}_1, \dots, \tilde{\omega}_l), \\ \varphi &= (\varphi_1, \dots, \varphi_h), \quad \tilde{\varphi} = (\tilde{\varphi}_1, \dots, \tilde{\varphi}_h); \end{aligned}$$

$$\omega = (\omega_1, \dots, \omega_l, \tilde{\omega}_1, \dots, \tilde{\omega}_l), \quad \varphi = (\varphi_1, \dots, \varphi_h, \tilde{\varphi}_1, \dots, \tilde{\varphi}_h).$$

$$\begin{aligned} \omega &= (\omega_1, \dots, \omega_l), \quad \tilde{\omega} = (\tilde{\omega}_1, \dots, \tilde{\omega}_l), \\ \varphi &= (\varphi_1, \dots, \varphi_h), \quad \tilde{\varphi} = (\tilde{\varphi}_1, \dots, \tilde{\varphi}_h); \end{aligned}$$

$$\begin{aligned} F_1^1(i) &= \langle \omega, \mathbf{V}_i \rangle, \quad F_2^1(i, j) = \langle \tilde{\omega}, \mathbf{E}_{ij} \rangle, \\ F_1^2(i) &= \langle \varphi, \mathbf{V}_i \rangle, \quad F_2^2(i, j) = \langle \tilde{\varphi}, \mathbf{E}_{ij} \rangle. \end{aligned}$$

$F_1^k(i), F_2^k(i, j), k \in \{1, 2\}$, — начальные значения весов.

$$\omega = (\omega_1, \dots, \omega_l, \tilde{\omega}_1, \dots, \tilde{\omega}_l), \quad \varphi = (\varphi_1, \dots, \varphi_h, \tilde{\varphi}_1, \dots, \tilde{\varphi}_h).$$

$$\omega = (\omega_1, \dots, \omega_l), \quad \tilde{\omega} = (\tilde{\omega}_1, \dots, \tilde{\omega}_l), \\ \varphi = (\varphi_1, \dots, \varphi_h), \quad \tilde{\varphi} = (\tilde{\varphi}_1, \dots, \tilde{\varphi}_h);$$

$$F_1^1(i) = \langle \omega, \mathbf{V}_i \rangle, \quad F_2^1(i, j) = \langle \tilde{\omega}, \mathbf{E}_{ij} \rangle, \\ F_1^2(i) = \langle \varphi, \mathbf{V}_i \rangle, \quad F_2^2(i, j) = \langle \tilde{\varphi}, \mathbf{E}_{ij} \rangle.$$

$F_1^k(i), F_2^k(i, j), k \in \{1, 2\}$, — начальные значения весов.

$$G^k(i) = \mu \frac{F_1^k(i)}{\sum_{j \in V} F_1^k(j)} + (1 - \mu) \sum_{\tilde{i}: \tilde{i} \rightarrow i \in E} \frac{F_2^k(\tilde{i}, i)}{\sum_{j: \tilde{i} \rightarrow j \in E} F_2^k(\tilde{i}, j)} G^k(\tilde{i}),$$

$G^1(i), G^2(i)$ — окончательные значения весов.

$$f(\omega, i) = G^1(i), \quad g(\varphi, i \rightarrow j) = G^2(j).$$

$$G_{i+1}^k = A_k G_i^k, \quad k \in \{1, 2\},$$

$$A_k = (a_{i,j}^k), \quad a_{i,j}^k = p_k(i \rightarrow j),$$

$$p_k(i \rightarrow j) = \mu \frac{F_1^k(j)}{\sum_{h \in V} F_1^k(h)} + (1 - \mu) \frac{F_2^k(i, j)}{\sum_{\tilde{j}: i \rightarrow \tilde{j} \in E} F_2^k(i, \tilde{j})}.$$

$$G_{i+1}^k = A_k G_i^k, \quad k \in \{1, 2\},$$

$$A_k = (a_{i,j}^k), \quad a_{i,j}^k = p_k(i \rightarrow j),$$

$$p_k(i \rightarrow j) = \mu \frac{F_1^k(j)}{\sum_{h \in V} F_1^k(h)} + (1 - \mu) \frac{F_2^k(i, j)}{\sum_{\tilde{j}: i \rightarrow \tilde{j} \in E} F_2^k(i, \tilde{j})}.$$

$$\pi_{i+1} = A\pi_i,$$

$$A = (a_{i,j}), \quad a_{i,j} = p(i \rightarrow j),$$

$$p(i \rightarrow j) = \alpha \frac{G^1(j)}{\sum_{h \in V} G^1(h)} + (1 - \alpha) \frac{G^2(j)}{\sum_{\tilde{j}: i \rightarrow \tilde{j} \in E} (G^2(\tilde{j}))}.$$

$$\pi_{i+1} = A\pi_i,$$

$$G_{i+1}^k = A_k G_i^k, \quad k \in \{1, 2\}.$$

$$\pi_{i+1} = A\pi_i,$$

$$G_{i+1}^k = A_k G_i^k, \quad k \in \{1, 2\}.$$

$$\begin{aligned} \left(\frac{\partial \pi}{\partial \chi} \right)_{i+1} &= \frac{\partial A}{\partial \chi} \pi_i + A \left(\frac{\pi}{\chi} \right)_i, \\ \left(\frac{\partial G^k}{\partial \chi} \right)_{i+1} &= \frac{\partial A_k}{\partial \chi} \pi_i + A \left(\frac{G^k}{\chi} \right)_i. \end{aligned}$$

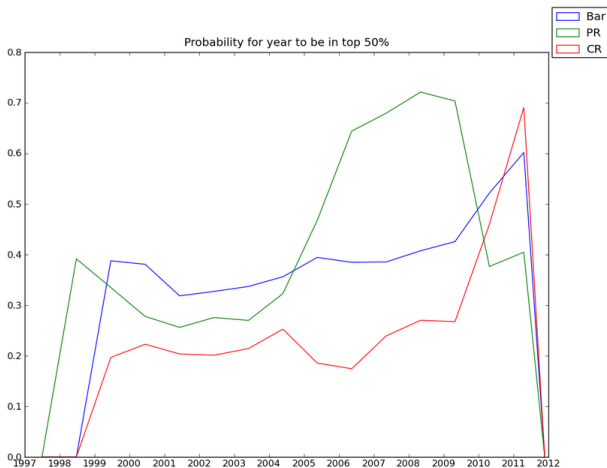
- Свежесть

- Свежесть
- Зависимость от запросов

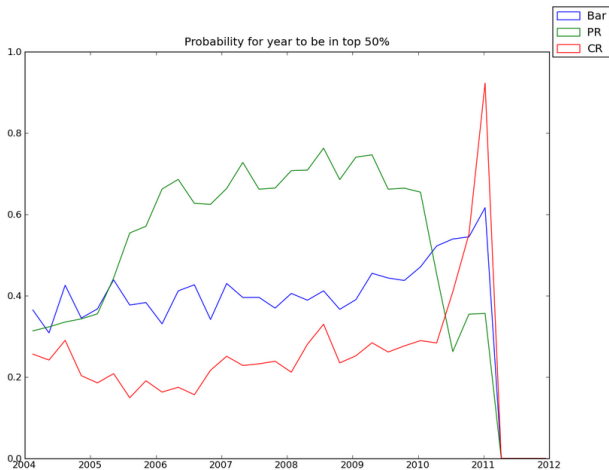
- Свежесть
- Зависимость от запросов
- Бинарные факторы

- Свежесть
- Зависимость от запросов
- Бинарные факторы
- Различные графы

Отрицательная корреляция classic PageRank с датой последней активности на странице [Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, Katsumi Tanaka, proc. of JCDL'07, pp. 107–116, 2007].



Отрицательная корреляция classic PageRank с датой последней активности на странице [Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, Katsumi Tanaka, proc. of JCDL'07, pp. 107–116, 2007].



- **Зависимость от in-degree страницы.** Значение фактора $w(p)$ растет с ростом количества ссылок на страницу p .

- **Зависимость от in-degree страницы.** Значение фактора $w(p)$ растет с ростом количества ссылок на страницу p .
- **Зависимость от свежести страницы.** Значение фактора $w(p)$ уменьшается по мере устаревания страницы p .

- **Зависимость от in-degree страницы.** Значение фактора $w(p)$ растет с ростом количества ссылок на страницу p .
- **Зависимость от свежести страницы.** Значение фактора $w(p)$ уменьшается по мере устаревания страницы p .
- **Зависимость от свежести ссылок.** При прочих равных условиях страница с меньшим количеством свежих входящих и исходящих ссылок должна получать меньшее $w(p)$.

- **Зависимость от in-degree страницы.** Значение фактора $w(p)$ растет с ростом количества ссылок на страницу p .
- **Зависимость от свежести страницы.** Значение фактора $w(p)$ уменьшается по мере устаревания страницы p .
- **Зависимость от свежести ссылок.** При прочих равных условиях страница с меньшим количеством свежих входящих и исходящих ссылок должна получать меньшее $w(p)$.
- **Зависимость от свежести соседей.** При прочих равных условиях страница с большим количеством устаревших соседей получает меньшее $w(p)$.

- Зависимость только от одного типа активности

- 1) P. S. Yu, X. Li, and B. Liu, Proc. WWW'04, pp. 448–449, 2004,
- 2) E. Amitay, D. Carmel, M. Herscovici, R. Lempel, and A. Soffer. Journal of the American Society for Information Science and Technology, 55(14), pp. 1270–1281, 2004.

- **Зависимость только от одного типа активности**

- 1) P. S. Yu, X. Li, and B. Liu, Proc. WWW'04, pp. 448–449, 2004,
- 2) E. Amitay, D. Carmel, M. Herscovici, R. Lempel, and A. Soffer. Journal of the American Society for Information Science and Technology, 55(14), pp. 1270–1281, 2004.

- **Отсутствие зависимости от активности соседей**

- 1) K. Berberich, M. Vazirgiannis, G. Weikum, Int. Math., 2(3), pp. 301–332, 2005.
- 2) K. Berberich, S. Bedathur, M. Vazirgiannis, G. Weikum. Proc. WWW'06, pp. 937–938, 2006.

- **Зависимость только от одного типа активности**

- 1) P. S. Yu, X. Li, and B. Liu, Proc. WWW'04, pp. 448–449, 2004,
- 2) E. Amitay, D. Carmel, M. Herscovici, R. Lempel, and A. Soffer. Journal of the American Society for Information Science and Technology, 55(14), pp. 1270–1281, 2004.

- **Отсутствие зависимости от активности соседей**

- 1) K. Berberich, M. Vazirgiannis, G. Weikum, Int. Math., 2(3), pp. 301–332, 2005.
- 2) K. Berberich, S. Bedathur, M. Vazirgiannis, G. Weikum. Proc. WWW'06, pp. 937–938, 2006.

- **Все 4 типа активности учтены**

Na Dai and Brian D. Davison, Proc. SIGIR'10, pp. 114–121, 2010.

Период времени $[\tau, T]$, количество срезов K ,
 $[t_{i-1}, t_i]$, $i \in \{1, \dots, K\}$, $t_0 = \tau$, $t_i - t_{i-1} = \frac{T-\tau}{K}$,
 G_i — срез графа в момент времени t_i , $N_i = v(G_i)$.

1) Актуальность $\Delta F_i(p)$ для графа G_i .

Activity Type	Features
Creation of the page p , $t(p) = i$	$n_1(i) ++$
Creation of a link $\tilde{p} \rightarrow p$, $t(\tilde{p} \rightarrow p) = i$ $\max(t(p), t(\tilde{p})) < i$	$n_2(i) ++$
$\max(t(p), t(\tilde{p})) = i$	$n_3(i) ++$
Creation of a link $p \rightarrow \tilde{p}$, $t(\tilde{p} \rightarrow p) = i$ $\max(t(p), t(\tilde{p})) < i$	$n_4(i) ++$
$\max(t(p), t(\tilde{p})) = i$	$n_5(i) ++$

2) Определение окончательного значения актуальности.

$$F_i(p) = \beta F_{i-1}(p) + \Delta F_i(p).$$

Период времени $[\tau, T]$, количество срезов K ,
 $[t_{i-1}, t_i]$, $i \in \{1, \dots, K\}$, $t_0 = \tau$, $t_i - t_{i-1} = \frac{T-\tau}{K}$,
 G_i — срез графа в момент времени t_i , $N_i = v(G_i)$.

1) Актуальность $\Delta F_i(p)$ для графа G_i .

Activity Type	Features
Creation of the page p , $t(p) = i$	$n_1(i) ++$
Creation of a link $\tilde{p} \rightarrow p$, $t(\tilde{p} \rightarrow p) = i$ $\max(t(p), t(\tilde{p})) < i$	$n_2(i) ++$
$\max(t(p), t(\tilde{p})) = i$	$n_3(i) ++$
Creation of a link $p \rightarrow \tilde{p}$, $t(\tilde{p} \rightarrow p) = i$ $\max(t(p), t(\tilde{p})) < i$	$n_4(i) ++$
$\max(t(p), t(\tilde{p})) = i$	$n_5(i) ++$

2) Определение окончательного значения актуальности.

$$F_i(p) = \beta F_{i-1}(p) + \Delta F_i(p).$$

$$K = 1 \Rightarrow \text{APR} = \text{NSSP}.$$

Parameter	Description
$[\tau, T]$	the considered period of time
K	the number of time intervals
a^1	the gain $F_1^1(p, i)$ receives if $t(p) = i$
b^1	the gain $F_2^1(p, i)$ receives if $t(p) = i$
c^1	the gain $F_1^2(p, i)$ receives if $t(p) = i$
d^1	the gain $F_2^2(p, i)$ receives if $t(p) = i$
a^2	the gain $F_1^1(p, i)$ receives if a user clicks to p in the i -th period
b^2	the gain $F_2^1(p, i)$ receives if a user clicks to p in the i -th period
c^2	the gain $F_1^2(p, i)$ receives if a user clicks to p in the i -th period
d^2	the gain $F_2^2(p, i)$ receives if a user clicks to p in the i -th period
a^3	the gain $F_1^1(p, i)$ receives if a user clicks to p in the i -th period
...	...
α, μ^1, μ^2	damping factors
β	the rate of decreasing of $F_i(p)$

Таблица : Parameters of Fresh BrowseRank.

Iteration	Method	NDCG@3,@5
First	SSP	0.90367, 0.92194
	NSSP	0.89904, 0.91828
Last	SSP	0.90377, 0.922
	NSSP	0.93267, 0.94588

Таблица : Performance of SSP and NSSP.

Method	NDCG@3,@5
SSP	0.8951, 0.9215
NSSP	0.8991, 0.9182

Таблица : Performance of the algorithms with $\alpha = 0.15$.