Использование истории интернет-поиска для улучшения его качества

Диссертация на соискание учёной степени магистра

Выполнил: студент группы 6057/2 Фишков А. А.
Научный руководитель: к. ф.-м. н. Николенко С. И.

Введение

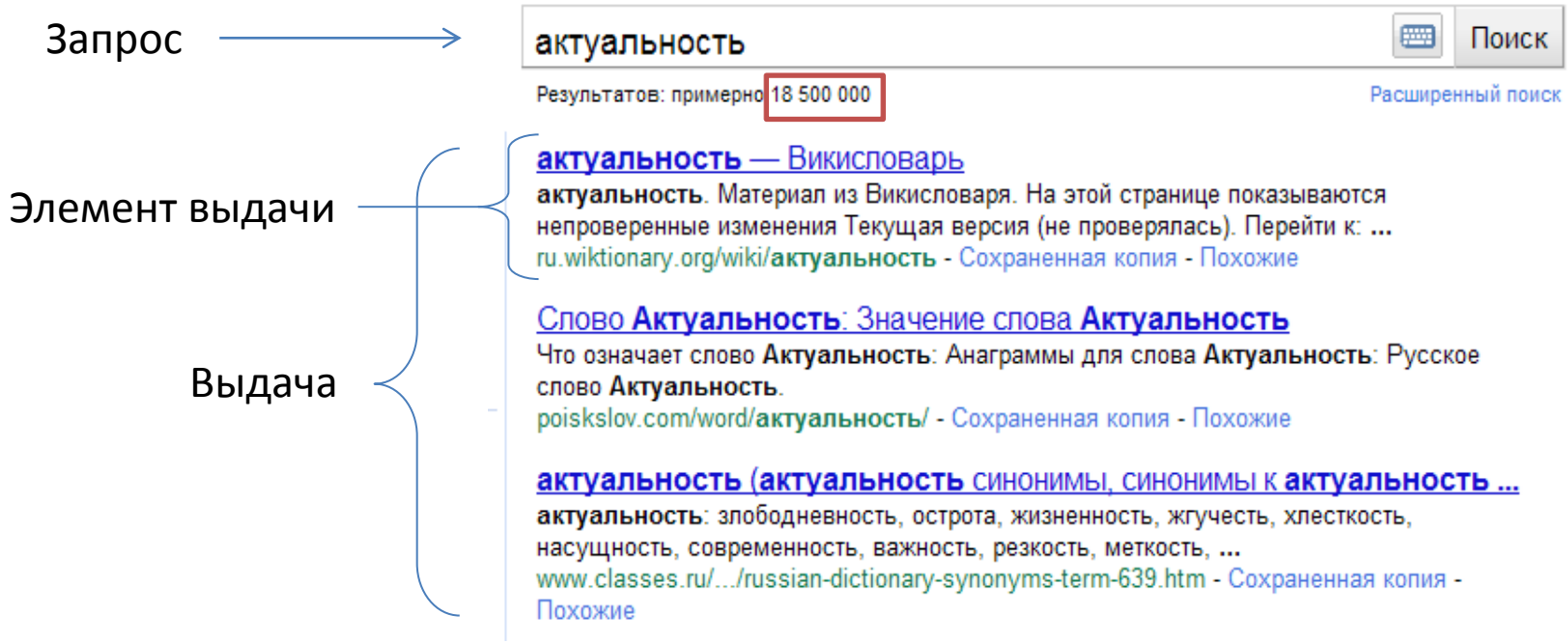


Рисунок 1: поисковая система Google

- Релевантность – степень соответствия документа запросу
- Ранжирование – упорядочивание документов в выдаче по релевантности

Введение

- Экспертные оценки релевантности
  - дороги
  - могут быть неточными или устаревать
  - доступны в ограниченном количестве
- Истории пользовательских действий
  - бесплатны
  - актуальны
  - доступны в больших количествах

Цели работы

- Разработать модель поведения пользователя для получения оценок релевантности документов по истории пользовательских действий
- Исследовать применение полученных оценок для построения ранжирующей функции

Исходные данные

- Предоставлены компанией Яндекс в рамках конкурса «Интернет-математика»
- Полностью анонимизированы
- Отрывок истории:

```
109 0   Q 103 2 67 218 384 18 45 514 414 315 143 499
109 103 C 218
109 955 Q 100 2 515 539 11 179 585 162 15 162 585 218
109 960 C 539
```

Сессия

Запрос

Клик

Выдача

Существующие модели

Название	Position bias	Presentation bias	Intent bias	Несколько запросов	Скорость работы
CCM	+	-	-	-	+
DBN	+	+	-	-	+
TCM	-	+	+	+	-
SCM	**+**	**+**	**+**	**+**	**+**

Таблица 1: Существующие модели

- пользователь неточно формулирует запрос к поисковой системе (intent bias)
- прежде чем сделать клик, пользователь просматривает описание элемента выдачи (presentation bias)
- пользователь просматривает выдачу строго последовательно и покидает поиск, как только находит первый устраивающий его документ (position bias)
- пользователь с меньшей вероятностью будет кликать на ранее показанный документ

Байесовские сети

- Способ графически представить факторизацию распределения

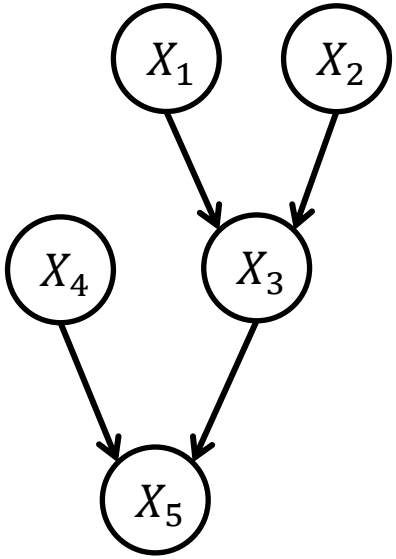


Рисунок 2: Пример байесовской сети

$$P(X_1, X_2, X_3, X_4, X_5) = P(X_1)P(X_2)P(X_3 \mid X_1, X_2)P(X_4)P(X_5 \mid X_3, X_4)$$

Новая модель поведения пользователя

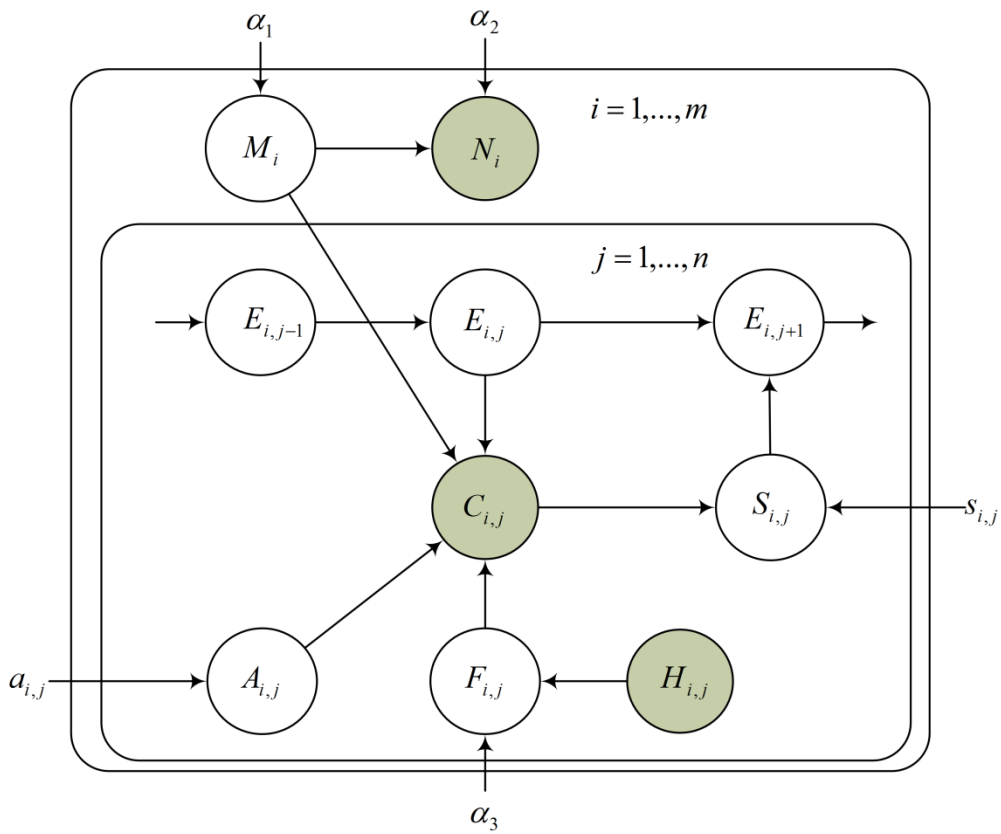


Рисунок 4: модель SCM

$E_{i,j}$	Просмотр описания
$A_{i,j}$	Привлекает внимание
$C_{i,j}$	Клик
$S_{i,j}$	Подходящий документ
$F_{i,j}$	Свежий документ
$H_{i,j}$	Был показан ранее
M_i	Выдача соответствует запросу
N_i	Продолжение поиска
α_1	$P(M_i = 1)$
α_2	$P(M_i = 1 \mid N_i = 1)$
α_3	$P(F_{i,j} = 1 \mid H_{i,j} = 1)$
$a_{i,j}$	$P(A_{i,j} = 1)$
$s_{i,j}$	$P(S_{i,j} = 1 \mid C_{i,j} = 1)$
$a_{i,j}s_{i,j}$	Релевантность

Байесовский подход

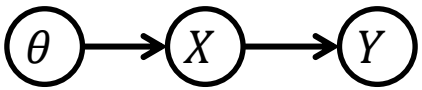


Рисунок 2: байесовская сеть

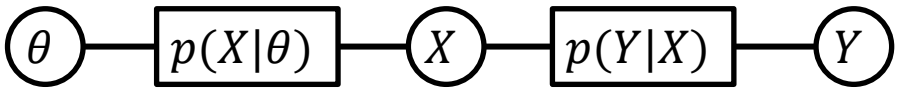


Рисунок 3: фактор-граф

$$p(\theta, X, Y) = p(\theta)p(X, Y|\theta) = p(\theta)p(X|\theta)p(Y|X)$$

$$p(\theta|Y = y_0) \propto p(Y = y_0|\theta)p(\theta) = \int p(X, Y = y_0|\theta)p(\theta)dx = \int p(\theta, X, Y = y_0)dx$$

- Точно – многомерное интегрирование
- Приближённо – передача сообщений на фактор-графе
- Expectation propagation[1] – маргинальные распределения из экспоненциального семейства
- В отличии от EM данные обрабатываются по мере поступления

[1]T. Minka, Expectation propagation for approximate Bayesian inference (UAI 2001)

Результаты сравнения моделей

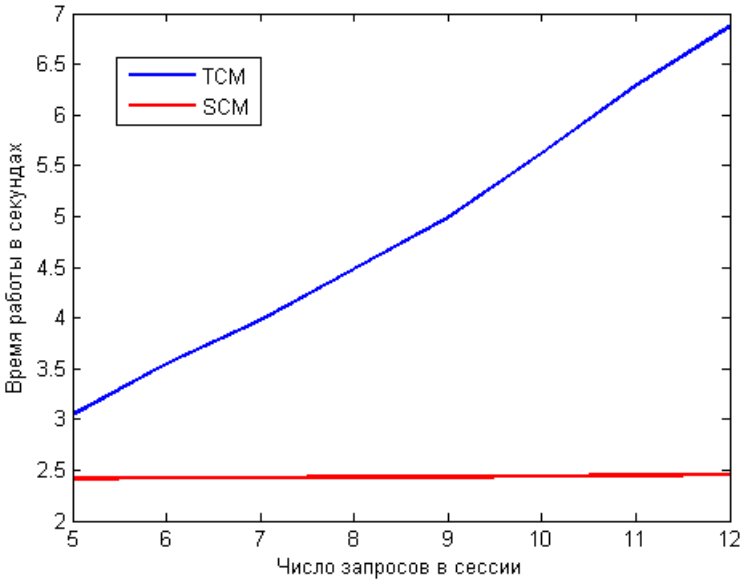


Рисунок 5: Сравнение скорости работы

Модель	AUC
CCM	0,5466
DBN	0,5654
TCM	0,5987
SCM	**0,6034**

Таблица 2: Ранжирование на основе релевантности

$$\widehat{AUC} = \frac{\sum r_i - n_+(n_+ + 1)/2}{n_+ n_-}$$

	-	-	-	-	+	-	+	+	+	+	AUC
i					1		2	3	4	5	24/25
r_i					5		7	8	9	10	

	1	2	3	4	5	6	7	8	9	10	AUC
№1	-	-	-	-	+	-	+	+	+	+	24/25
№2	+	-	-	-	-	+	+	+	+	-	16/25

Оценки релевантности как признаки для ранжирования

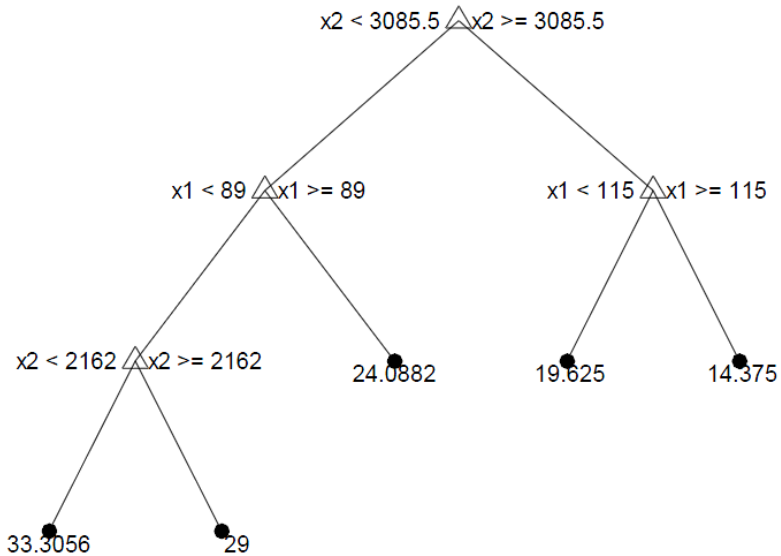


Рисунок 6: Пример дерева решений

$$h_{k+1}(x) = h_k(x) + tree(\{(x_i, y_i - h_k(x_i))\}_{i=1..l})$$

- Использовать дополнительные признаки документа и запроса из предоставленных данных
- Ранжирование по нескольким признакам
  - Деревья решений
  - Boosting[1] (усиление)

Модель	AUC
Ранжирование Яндекса	0,6119
CCM	0.6239
DBN	0.6252
TCM	0.6271
SCM	**0.6296**
Все модели и признаки вместе	**0.6315**

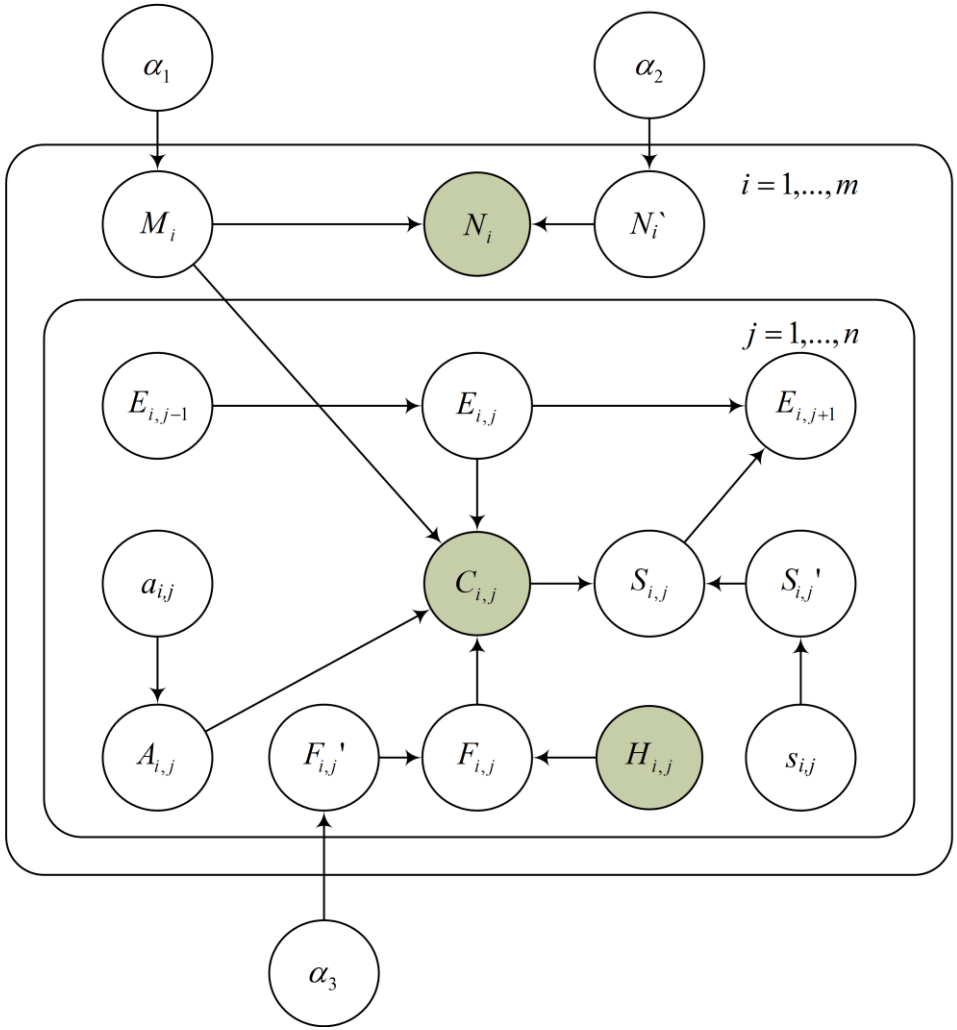
Таблица 3: Ранжирование на основе методов машинного обучения

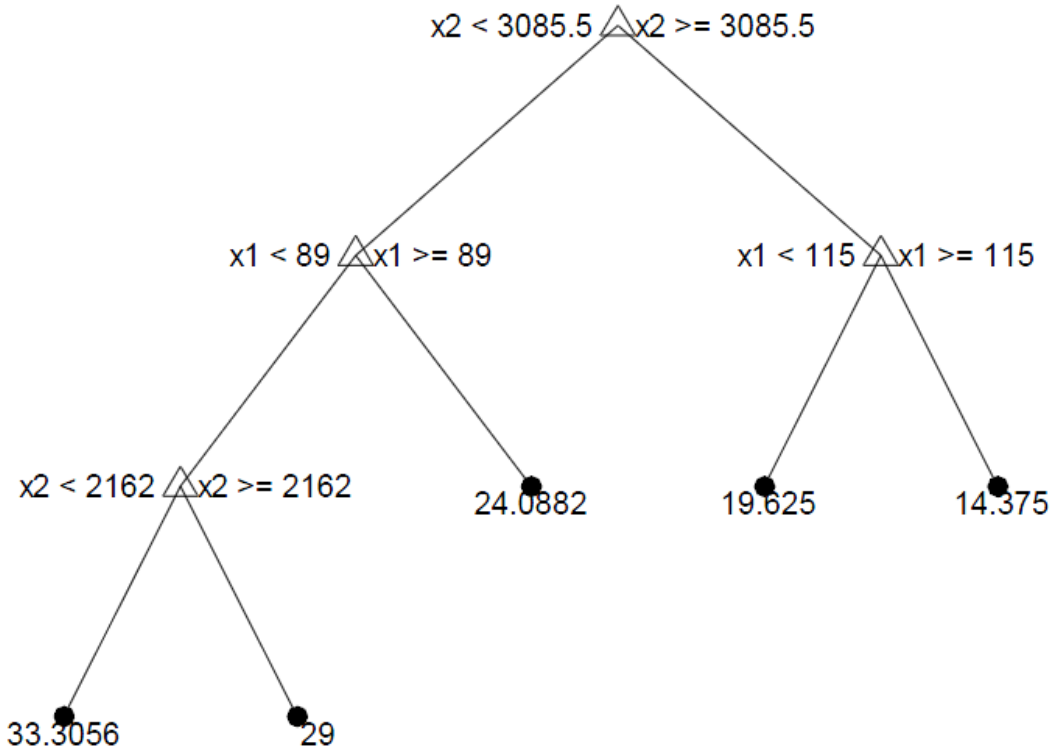
[1]J. H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, Vol. 29, No. 5, 2001

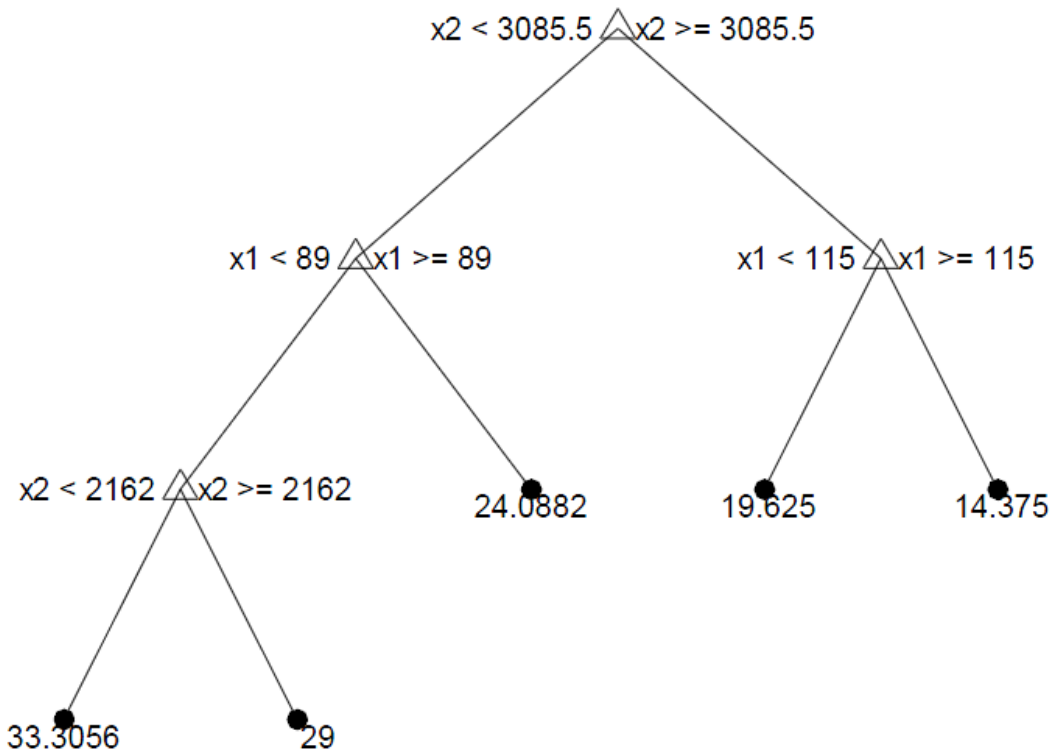
Заключение

- Предложена новая модель поведения пользователя SCM (session click model), учитывающая многократные запросы пользователя к поисковой системе, улучшающая существующие модели
- С использованием оценок релевантности документов, полученных с помощью предлагаемой модели, было получено улучшение качества поиска на тестовом множестве
- Результаты работы были опубликованы
  - EEML 2012, an ICFCA 2012 workshop
  - Труды СПИИРАН

Спасибо за внимание!

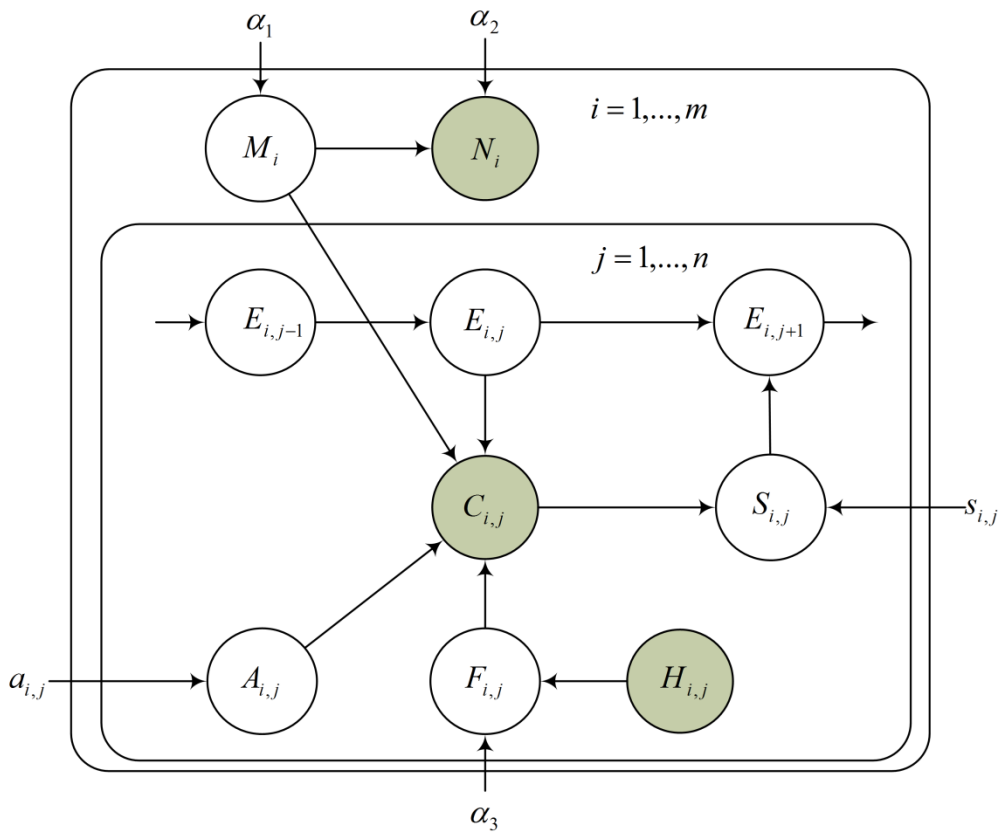

α_1
α_2
$i = 1,\ldots,m$
M_i
N_i
N_i`
$j = 1,\ldots,n$
$E_{i,j-1}$
$E_{i,j}$
$E_{i,j+1}$
$a_{i,j}$
$C_{i,j}$
$S_{i,j}$
$S_{i,j}$'
$A_{i,j}$
$F_{i,j}$'
$F_{i,j}$
$H_{i,j}$
$s_{i,j}$
α_3


x2 < 3085.5
x2 >= 3085.5
x1 < 89
x1 >= 89
x1 < 115
x1 >= 115
x2 < 2162
x2 >= 2162
24.0882
19.625
14.375
33.3056
29



$$h_{k+1}(x) = h_k(x) + tree(\{(x_i, y_i - h_k(x_i))\}_{i=1..l})$$

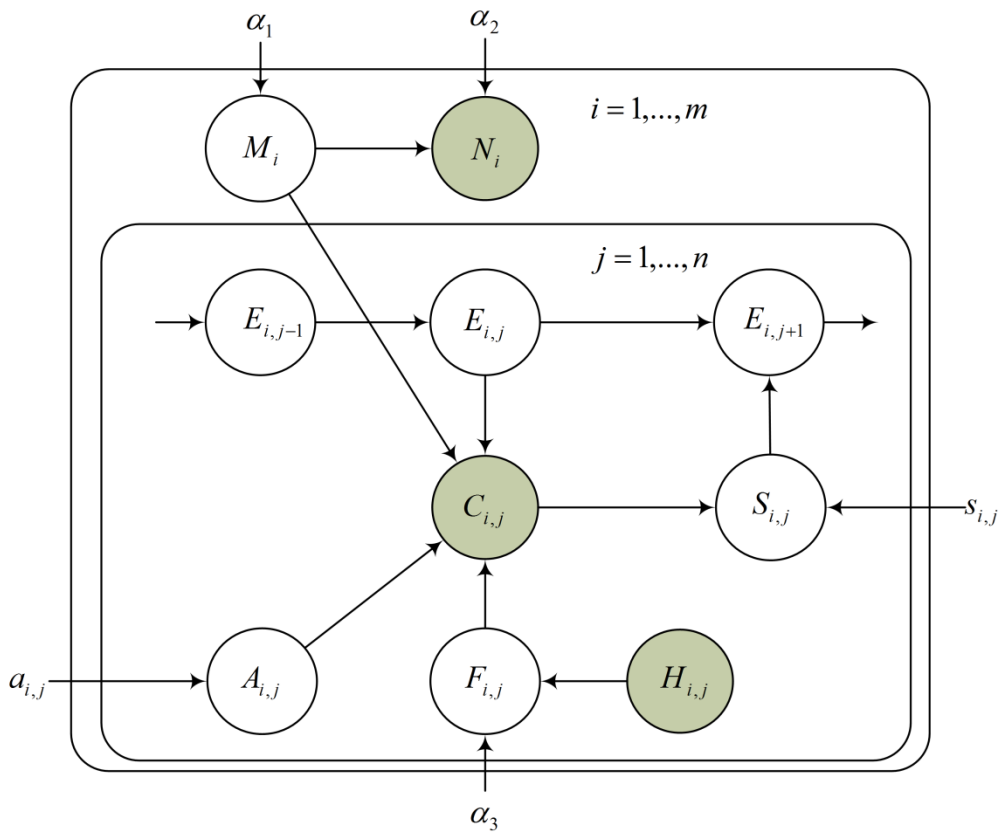
Новая модель поведения пользователя



$$r_{i,j} = P(S_{i,j} = 1 \mid E_{i,j} = 1, M_i = 1, F_{i,j} = 1) =$$
$$= P(S_{i,j} = 1 \mid C_{i,j} = 1) P(A_{i,j} = 1) = a_{i,j} s_{i,j}$$

Рисунок 4: модель SCM

Новая модель поведения пользователя



$$P(M_i = 1) = \alpha_1$$

$$P(N_i = 1 \mid M_i = 1) = \alpha_2$$

$$P(F_{i,j} = 1 \mid H_{i,j} = 0) = 1$$

$$P(F_{i,j} = 1 \mid H_{i,j} = 1) = \alpha_3$$

$$P(N_i = 1 \mid M_i = 0) = 1$$

$$P(A_{i,j} = 1) = a_{i,j}$$

$$E_{i,j-1} = 0 \Rightarrow E_{i,j} = 0$$

$$S_{i,j-1} = 1 \Rightarrow E_{i,j} = 0$$

$$E_{i,j} = 1 \Rightarrow E_{i,j-1} = 1, S_{i,j-1} = 0$$

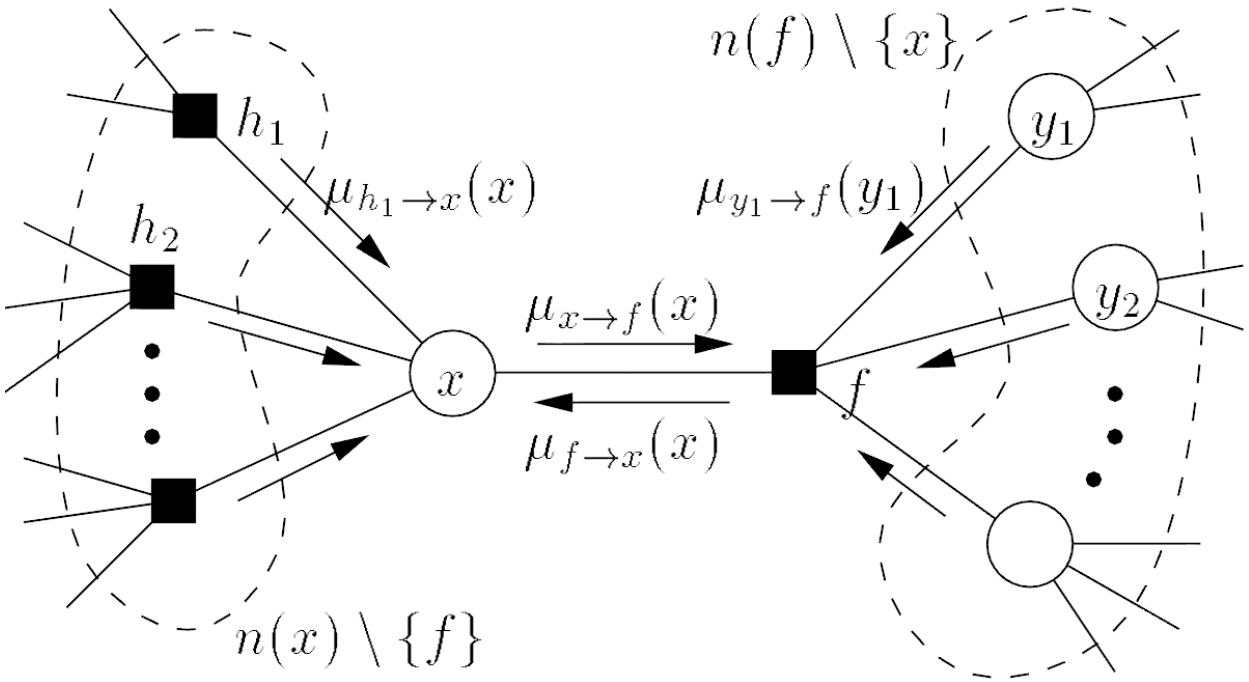
$$P(S_{i,j} = 1 \mid C_{i,j} = 1) = s_{i,j}$$

$$P(S_{i,j} = 1 \mid C_{i,j} = 0) = 0$$

$$C_{i,j} = 1 \Leftrightarrow M_i = 1, E_{i,j} = 1, S_{i,j} = 1, F_{i,j} = 1$$

Рисунок 4: модель SCM

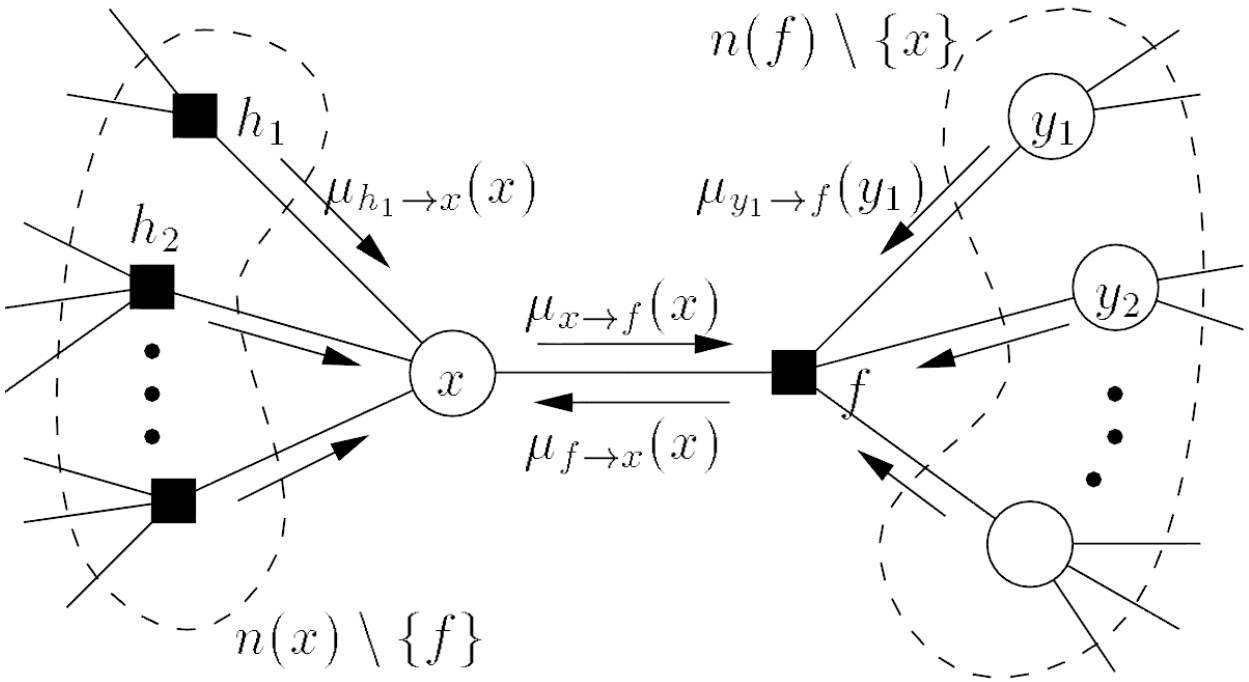
Фактор-граф



$$\mu_{x \to f}(x) = \prod_{i \in n(x) \setminus \{f\}} \mu_{h_i \to f}(x) \qquad \mu_{f \to x}(x) = \iiint f(x, y_1, \dots) \prod_{i \in n(f) \setminus \{x\}} \mu_{y_i \to f}(y_i)$$

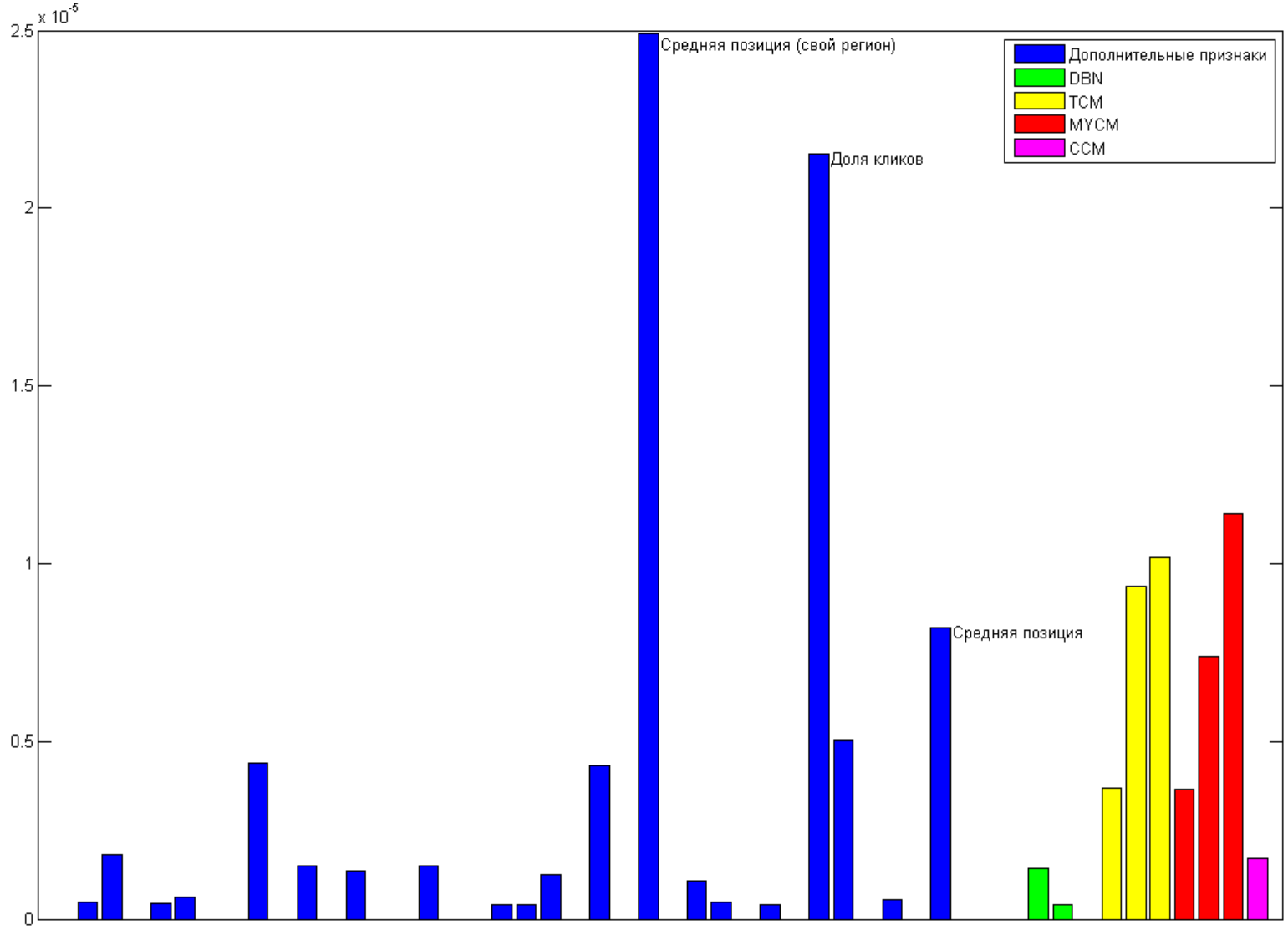
F. Kschischang, B. Frey. Factor graphs and the sum-product algorithm // IEEE Transactions on Information Theory, IEEE, 2001. P. 498–519

Фактор-граф



$$\mu_{x \to f}(x) = \prod_{i \in n(x) \setminus \{f\}} \mu_{h_i \to f}(x) \qquad \mu_{f \to x}(x) = \iiint f(x, y_1, \dots) \prod_{i \in n(f) \setminus \{x\}} \mu_{y_i \to f}(y_i)$$

$$\mu_{f \to x}(x) = \frac{1}{\mu_{x \to f}(x)} Proj \left[\mu_{x \to f}(x) \iiint f(x, y_1, \dots) \prod_{i \in n(f) \setminus \{x\}} \mu_{y_i \to f}(y_i) \right]$$


x 10^-5
2.5
2
1.5
1
0.5
0
Средняя позиция (свой регион)
Доля кликов
Средняя позиция
Дополнительные признаки
DBN
TCM
MYCM
CCM