

# **МЕТОД СКОЛЬЗЯЩЕГО КОНТРОЛЯ ДЛЯ ОЦЕНКИ КАЧЕСТВА РЕКОМЕНДАТЕЛЬНЫХ ИНТЕРНЕТ-СЕРВИСОВ**

**А.Ю. Каминская, Р.А. Магизов**

**Научный руководитель – Д.И. Игнатов**

**Государственный университет – Высшая школа экономики,  
Москва, Россия**

**2010**

# Содержание

- Постановка задачи
- Мотивация
- Примеры моделей рекомендательных систем
  - User-based RS
  - Item-based RS
- Выбор меры (метрики) сходства
- Методика сравнения
- Данные MovieLens и Yahoo
- Результаты
- Выводы и дальнейшая работа

# Постановка задачи

- Зная предпочтения конкретного пользователя и остальных, выдать ему релевантную рекомендацию
- Оценка (проверка) качества рекомендаций

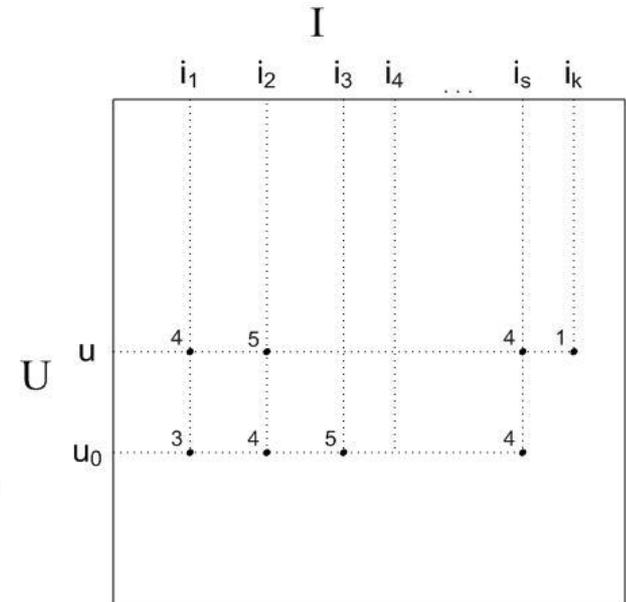
# Мотивация

- Существует огромное количество РС
- Требуются эффективные алгоритмы
  - Время выполнения
  - Качество рекомендаций
- Количество ошибок и полнота рекомендаций

# User-based RS

- целевой пользователь  $u_0$
- $u_0^I$  - предметы, которые он оценивал
- $sim(u_0, u)$  - сходство с пользователем  $u$
- $N(u_0) = \{u \mid sim(u_0, u) \geq \theta\}$  - top- $n$  ближайших к нему соседей,  $n$  определяет  $\theta$
- $N(u_0 \mid i) = \{u \mid i \in u^I \ \& \ u \in N(u_0)\}$  - пользователи из  $N(u_0)$  которые оценили предмет  $i$
- $r_{ui}$  - оценка предмета  $i$  пользователем  $u$
- предсказанная системой оценка для целевого пользователя:

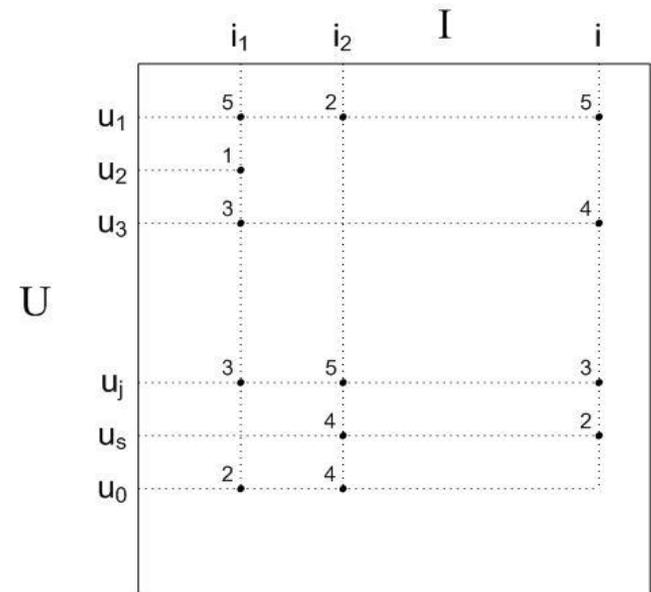
$$\hat{r}_{u_0 i} = \frac{\sum_{u \in N(u_0 \mid i)} sim(u_0, u) \times r_{ui}}{\sum_{u \in N(u_0 \mid i)} sim(u_0, u)}$$



# Item-based RS

- целевой пользователь  $u_0$
- $u_0^I$  - предметы, которые он оценивал
- $sim(i, j)$  - сходство товара  $i$  с товаром  $j$
- $N(i) = \{j \mid sim(i, j) \geq \theta\}$  - топ- $n$  ближайших к нему товаров, топ- $n$  определяет  $\theta$
- $N(i|u_0) = \{j \mid j \notin u_0^I \ \& \ i \in u_0^I \ \& \ j \in N(i)\}$  - для  $u_0$
- $r_{ui}$  - оценка предмета  $i$  пользователем  $u$
- предсказанная системой оценка для целевого пользователя:

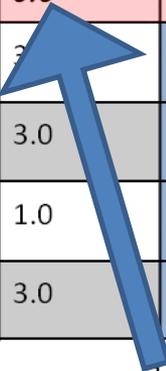
$$\hat{r}_{u_0 i} = \frac{\sum_{j \in N(i|u_0)} sim(i, j) \times r_{u_0 j}}{\sum_{j \in N(i|u_0)} sim(i, j)}$$



# Пример

## User-based RS

Пользователь	Сходство	«Мастер и Маргарита»	ВО «Мастер и Маргарита»	«Преступление и наказание»	ВО Преступление и наказание»	Рей Бредбери. Рассказы.	ВО Рей Бредбери. Рассказы.	Е. Гришковец. «Планета»	ВО «Планета»
V.M.		4.0		3.0		0.0		4.0	
SKY	0.5	4.0		4.0		3.0	$3.5 \times 0.5 = 1.75$	4.5	
OLA	0.5	4.5		4.0		3.0	$3.0 \times 0.5 = 1.5$	4.0	
GRY	0.94	4.0		2.5		1.0	$1.0 \times 0.94 = 0.94$	3.5	
IDI	0.87	2.0		2.5		3.0	$3.0 \times 0.87 = 2.6$	1.5	

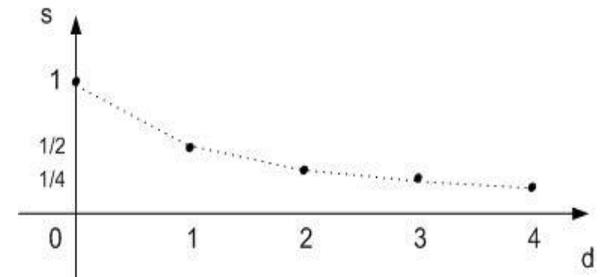

$$\frac{1.75 + 1.5 + 2.6 + 0.94}{0.5 + 0.5 + 0.87 + 0.94} = \frac{6.79}{2.81} = 2.4 \approx 2.5$$

# Выбор меры (метрики) сходства

- Сходство, основанное на расстоянии:

- Евклида  $d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$
- Хемминга  $d(x, y) = \sum_{x_i \neq y_i} 1$

$$\Rightarrow s = \frac{1}{1 + d}$$



- Корреляция как сходство:

- коэффициент Пирсона

- Косинусная мера  $\cos(x, y) = \frac{(x, y)}{|x| \cdot |y|}$

- Коэффициент Жаккара  $J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$

# Корреляция Пирсона

- $$pearson = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2 \cdot \sum_i (y_i - \bar{y})^2}$$

- Недостатки

- не определена на векторах с постоянными значениями:  $(4, 4, 4, \dots, 4)$

- теряются рекомендации

$$a = (0, 5, 5, 4)$$

$$b = (0, 4, 5, 0)$$

# Методика сравнения

- Метрики качества: точность и полнота рекомендаций
- Скользящий контроль (кросс-валидация)

# Точность и полнота

- Полнота – число релевантных рекомендаций к числу всех выбранных пользователем товаров

$$recall = \frac{|r_n(u) \cap u^I|}{|u^I|}$$

- Точность – число релевантных к числу всех рекомендаций

$$precision = \frac{|r_n(u) \cap u^I|}{|r_n(u)|}$$

# СКОЛЬЗЯЩИЙ КОНТРОЛЬ

- Разбиение на тестовую и обучающую выборки:

$$U = U_{training} \cup U_{test}, \quad U_{training} \cap U_{test} = \emptyset$$

- Соккрытие признаков для тестирования рекомендаций  $I_{hidden} \subseteq I$

- Вычисление точности и полноты для  $u \in U_{test}$  на признаках  $i \in I_{hidden}$  :

$$recall = \frac{|r_n(u) \cap u^I \cap I_{hidden}|}{|u^I \cap I_{hidden}|}$$

$$precision = \frac{|r_n(u) \cap u^I \cap I_{hidden}|}{|r_n(u) \cap I_{hidden}|}$$

	I <sub>visible</sub>	I <sub>hidden</sub>
U <sub>training</sub>	80%	
U <sub>test</sub>	20%	

# Точность и полнота: раскрытие неопределенностей

$$recall = \frac{|r_n(u) \cap u^I \cap I_{hidden}|}{|u^I \cap I_{hidden}|}$$

$$|u^I \cap I_{hidden}| = 0 \Rightarrow recall = 1$$

$$precision = \frac{|r_n(u) \cap u^I \cap I_{hidden}|}{|r_n(u) \cap I_{hidden}|}$$

$$|r_n(u) \cap I_{hidden}| = 0:$$

$$u^I = 0 \Rightarrow precision = 1$$

$$\text{else } precision = 0$$

# Алгоритм

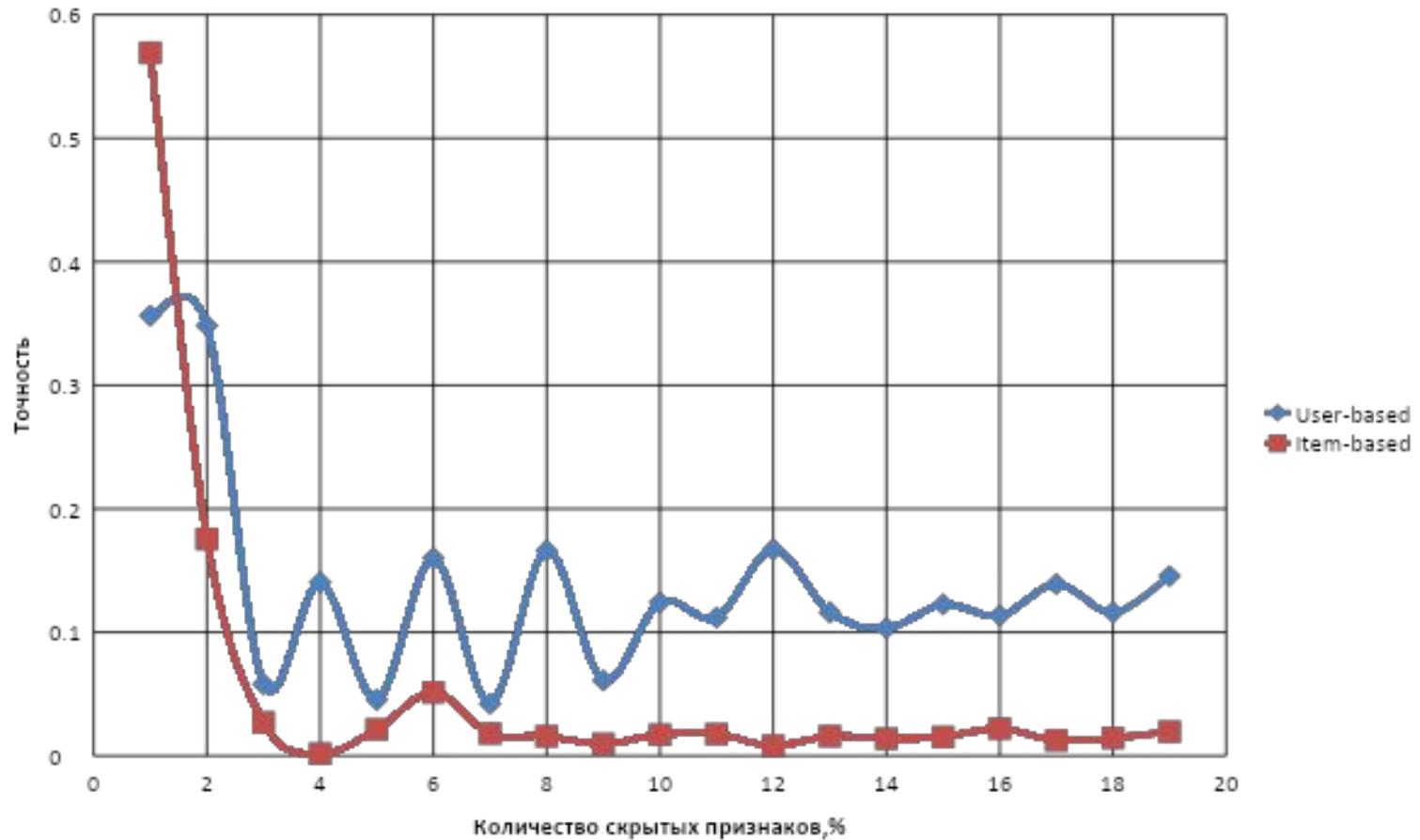
- **Параметры:**
  - test% - размер тестового множества
  - hidden% - размер скрытого множества признаков
  - p – число повторений разбиения на тестовое и обучающее множество
  - q – число повторений разбиения на оцененное и неоцененное множества признаков
- **Выход:** средние значения точности и полноты по множеству  $U_{\text{test}}$  и  $I_{\text{hidden}}$

# Данные MovieLens и Yahoo

- MovieLens 100K dataset:
  - 943 пользователя
  - 1,682 фильма
  - Каждый оценил как минимум 20 фильмов, всего 100,000 оценок
- Yahoo binary dataset:
  - 2,000 фирм
  - 3,000 рекламных словосочетаний
  - 92,345 ненулевых ячеек

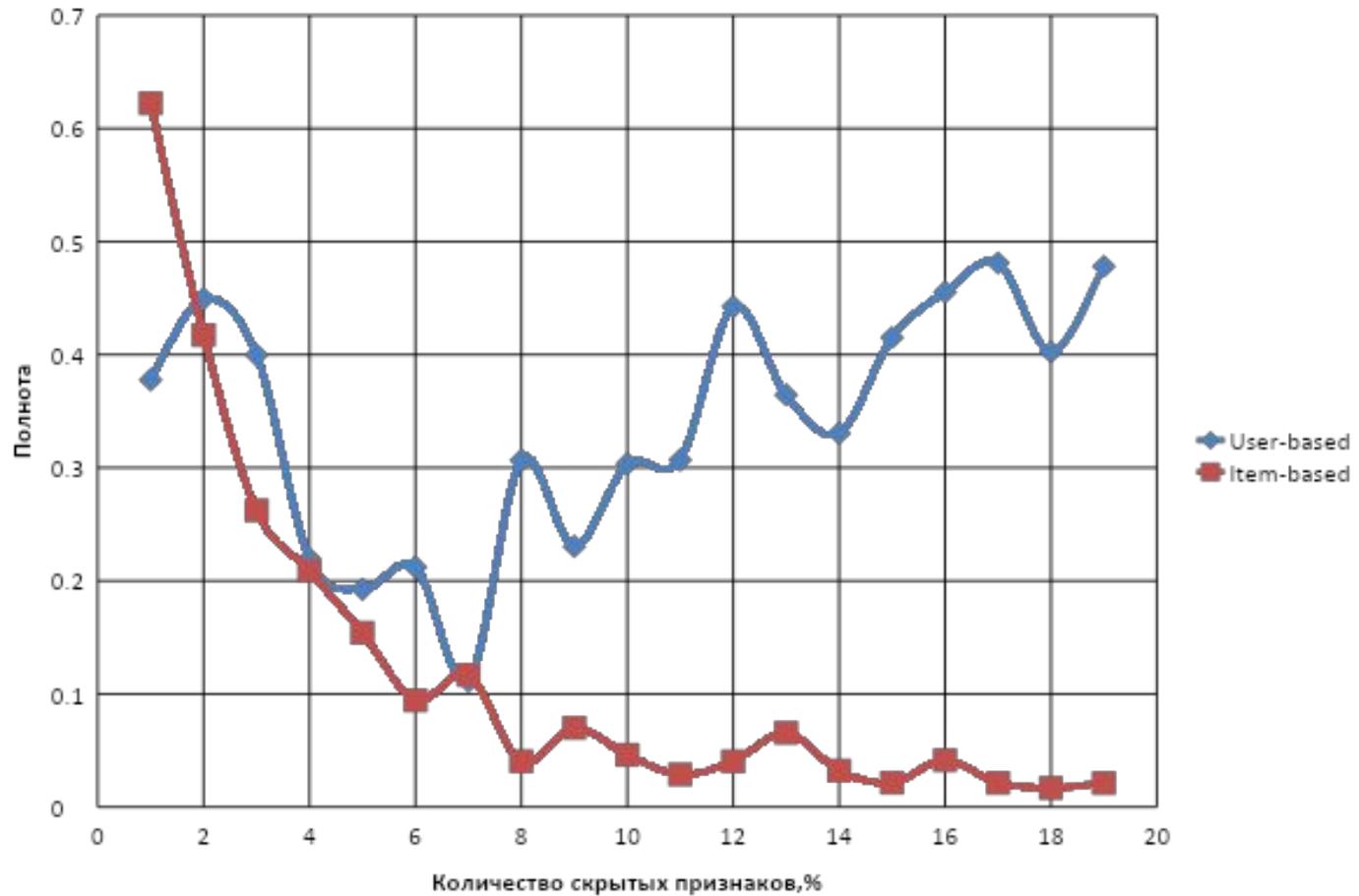
# Результаты

Зависимость точности от количества скрытых признаков



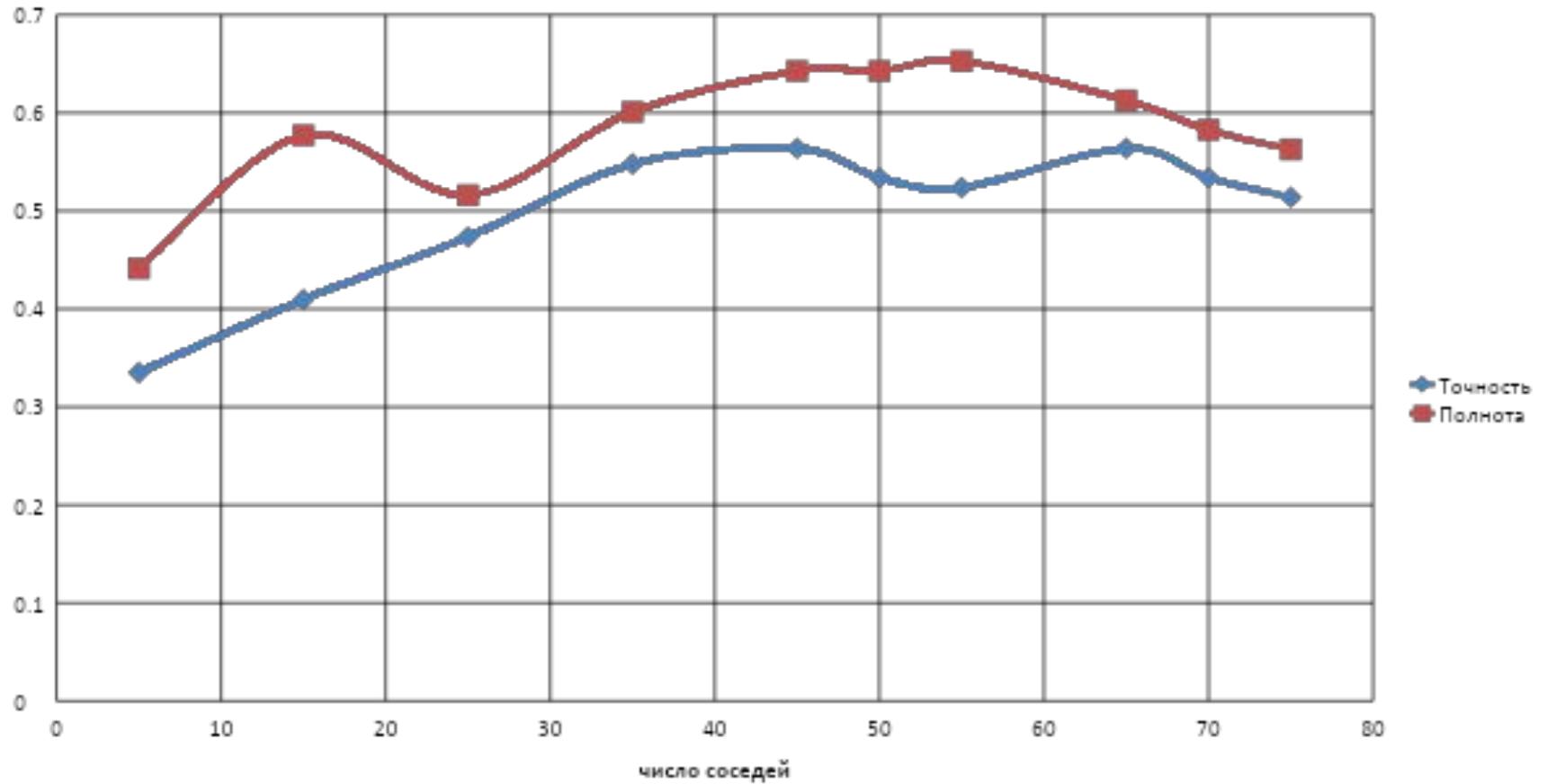
# Результаты

Зависимость полноты от количества скрытых признаков



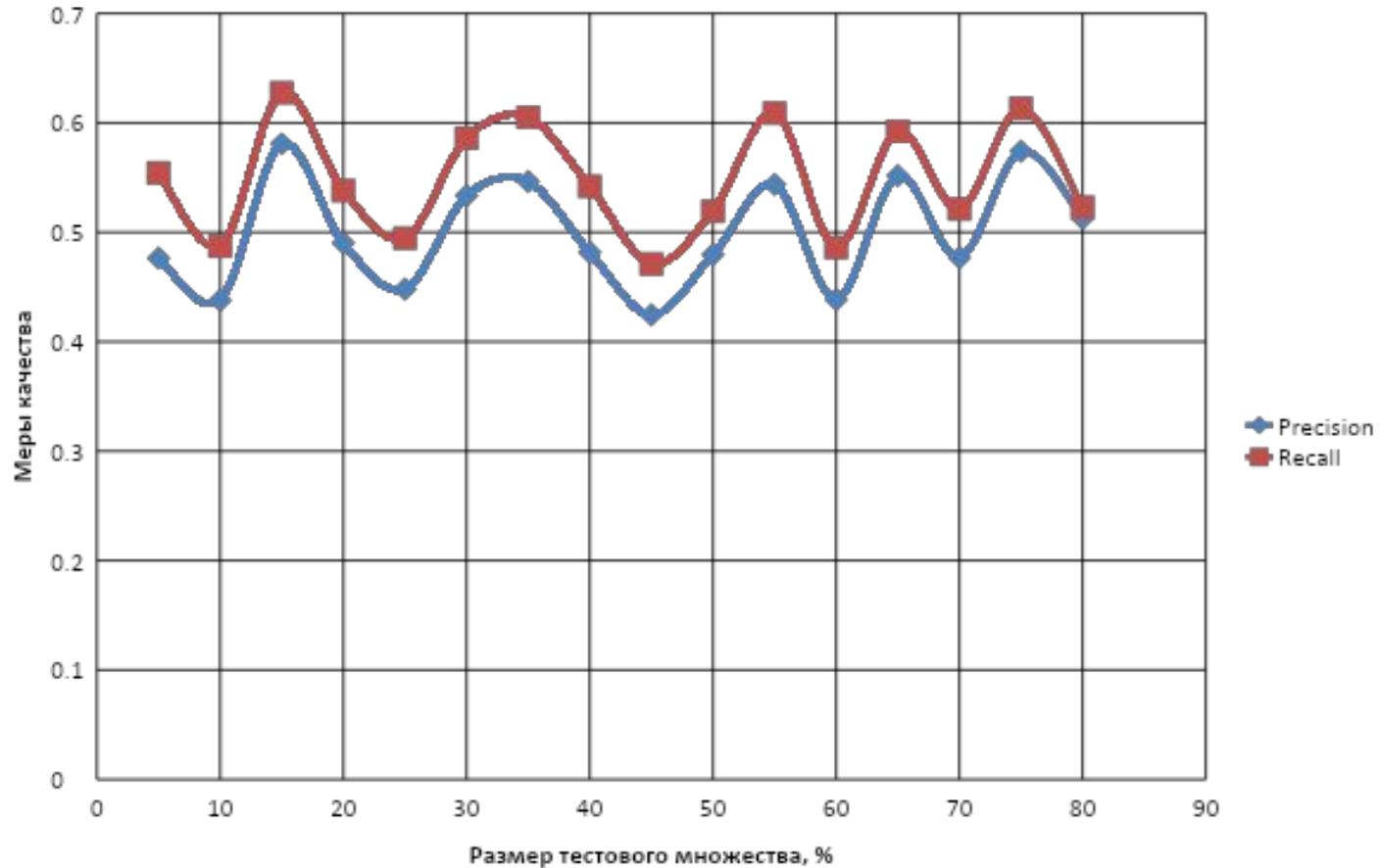
# Результаты

Зависимость точности и полноты от числа соседей



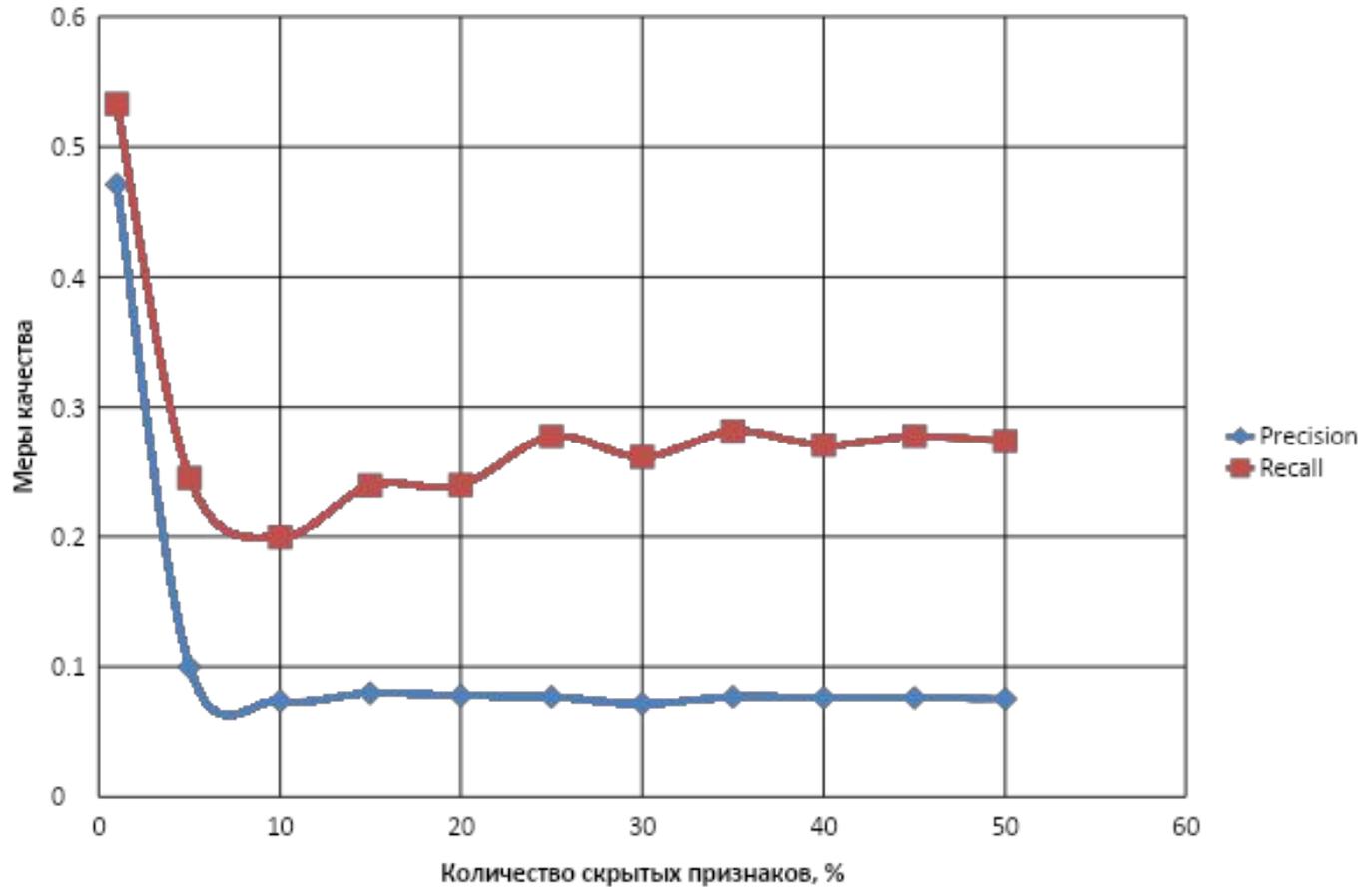
# Результаты

Зависимость точности и полноты от размера тестового множества для метода user-based



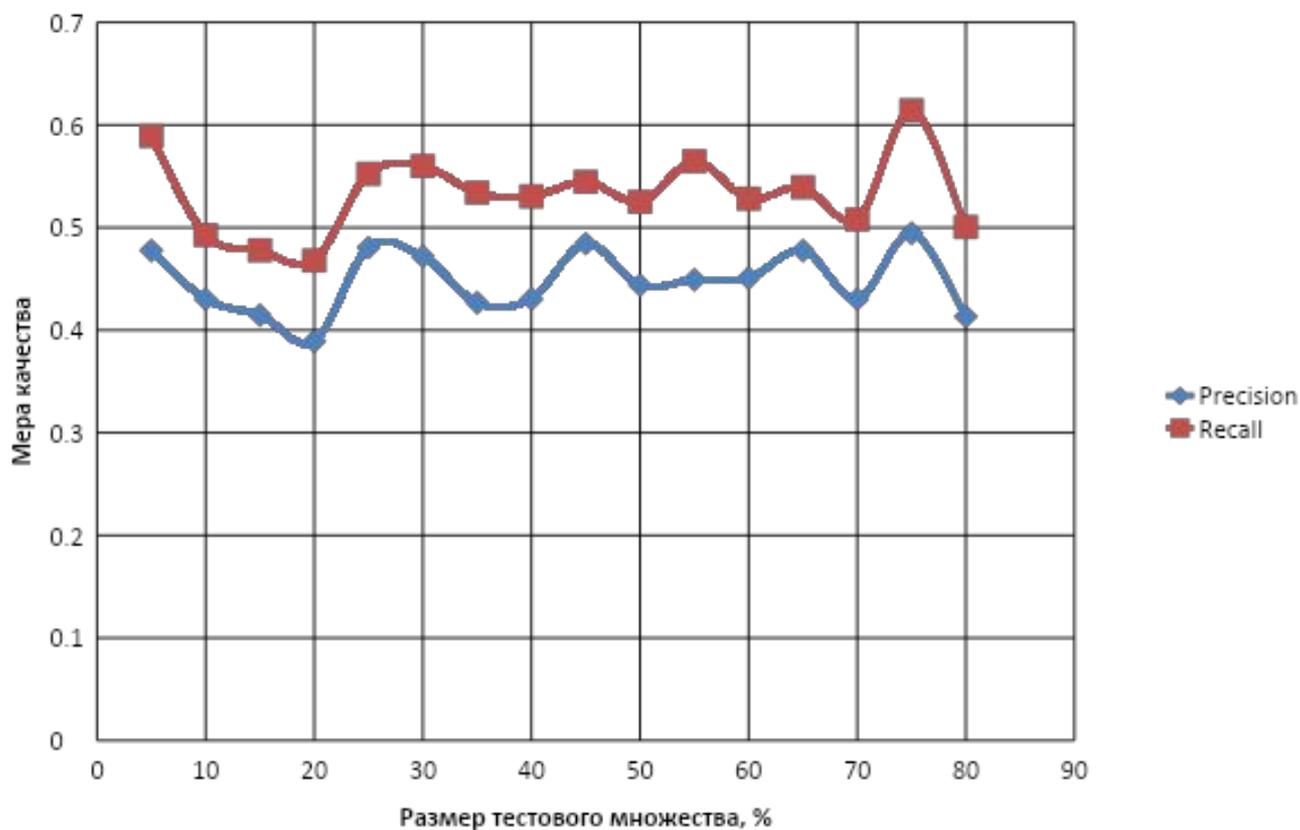
# Результаты

Зависимость точности и полноты от количества скрытых признаков для метода user-based



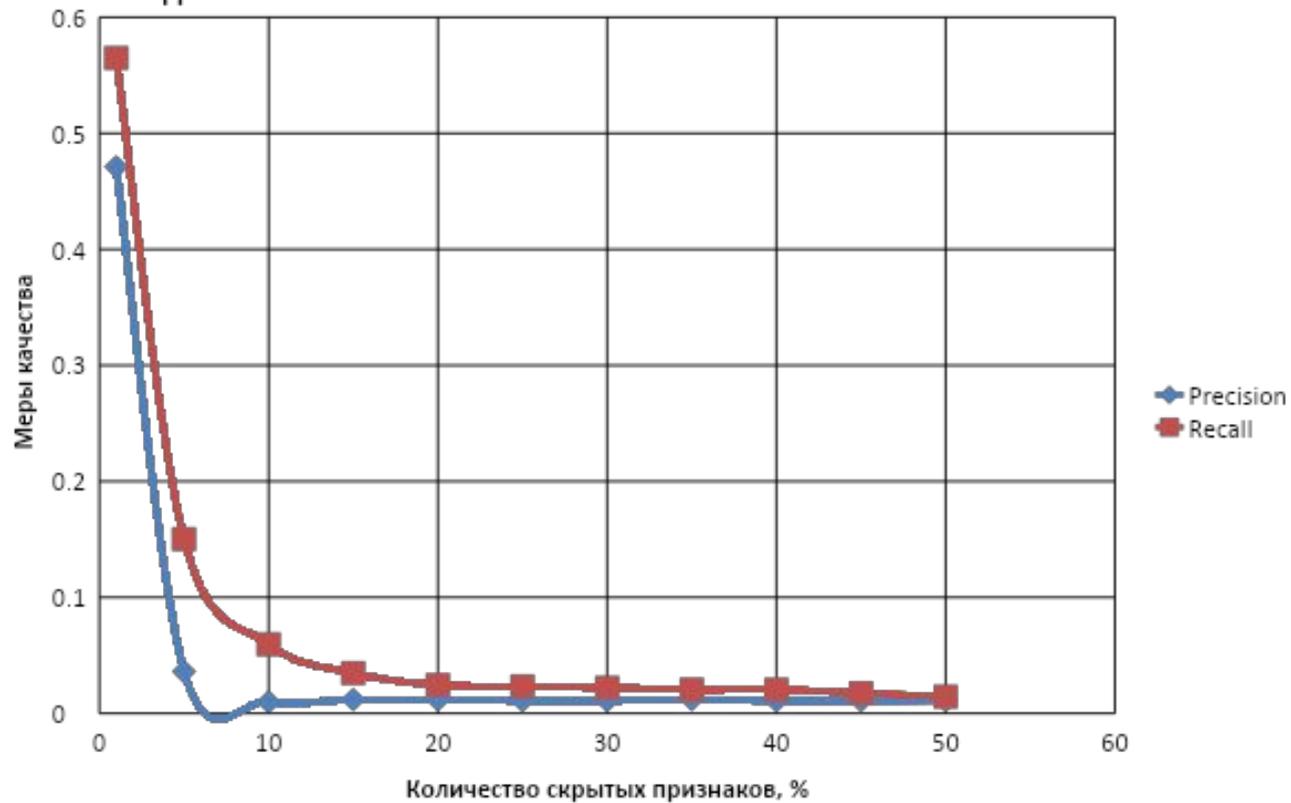
# Результаты

Зависимость точности и полноты от размера тестового множества для метода item-based



# Результаты

Зависимость точности и полноты от количества скрытых признаков для метода Item based



# Выводы и дальнейшая работа

- Предложенная методика позволяет оценить качество работы рекомендательной системы вне зависимости от выбора метода
- По-видимому, впервые в экспериментах исследуется точность и полнота в зависимости от количества скрытых признаков
- Для сравнения методов необходимо проведение аналогичных экспериментов для более совершенных моделей РС, например, основанных на бикластеризации

Спасибо за внимание!