Методы и алгоритмы метрического анализа клиентских сред

В. А. Лексин

Заочная аспирантура МФТИ, III год

Научный руководитель К. В. Воронцов

Научные результаты, полученные за предыдущие годы

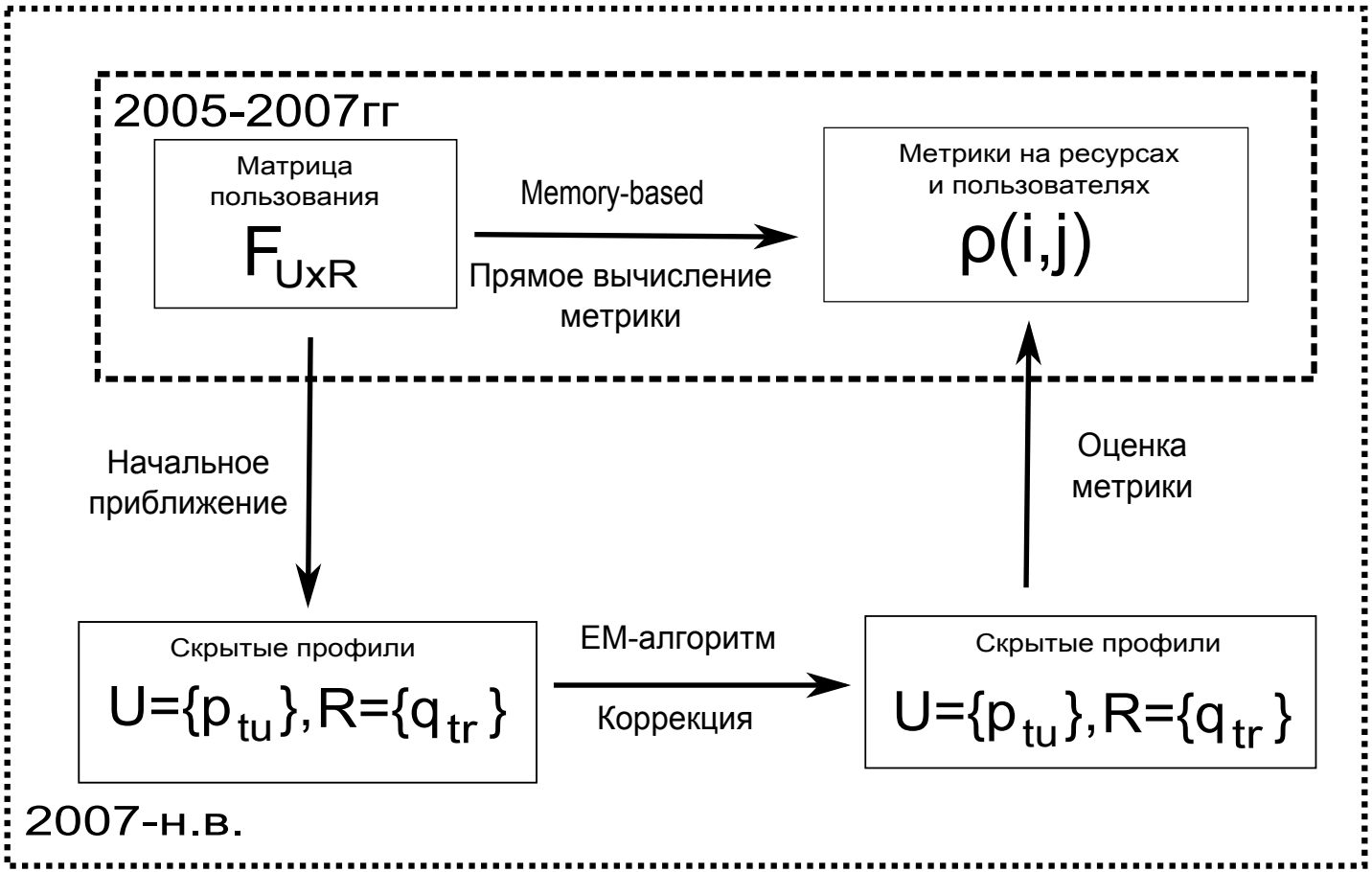

Рис.: Этапы исследований

Научные результаты, полученные за 2005-2007гг

- Исследованы методы прямого вычисления метрик в системах взаимодействующих объектов.
- Предложен вероятностный алгоритм, основанный на гипотезе о независимости посещения пользователем различных ресурсов (точный тест Фишера).
- Использована технология многомерного шкалирования и построения карт сходства для визуальной оценки кластерных свойств полученных метрик.
- Проведено исследование в рамках научной стипендии Яндекс: анализ лога поисковых запросов Яндекса методами АКС.
- Построены различные функционалы качества полученных метрик, основанные на частичной экспертной разметке ресурсов.

Научные результаты, полученные за 2007-2009гг

- Предложен двухступенчатый итерационный алгоритм, позволяющий оценивать скрытые профили малой размерности, а затем получать из этих профилей метрики на множествах пользователей и ресурсов.
- Произведено сравнение метода с прямым вычислением метрики, а также с другими известными подходами в коллаборативной фильтрации. Был выявлен ряд преимуществ рассматриваемого метода.
- Исследованы свойства переобучения алгоритма при слишком большом количестве итераций в EM-алгоритме.
- Предложена технология генерации модельных данных для оценки качества и настройки параметров алгоритма.

Научные результаты, полученные за последний год (2009-2010)

- Подключены данные о продажах товаров в крупной мебельной компании и данные форума.
- Исследованы алгоритмы построения ***иерархических профилей***. Предложен ***критерий достаточности*** наблюдений для определения компонент профилей (публикация в сборнике трудов МФТИ).
- Теоретически и экспериментально проверена взаимосвязь алгоритма с методом градиентного спуска. Получена ***оценка скорости сходимости*** EM-итераций (осталось несколько открытых вопросов).
- Теоретически доказана взаимосвязь итерационного процесса максимизации правдоподобия в pLSA с минимизацией дивергенции Кульбака-Лейблера, для которой уже доказана ***гарантированная сходимость к локальному экстремуму*** (осталось несколько открытых вопросов).
- Экспериментально изучены свойства влияния начального приближения профилей на итоговый результат.

Публикации 2005-2007

2007

- Оценивание сходства пользователей и ресурсов путем выявления скрытых тематических профилей. Тезисы для конференции МФТИ.
- Анализ клиентских сред: выявление скрытых профилей и оценивание сходства клиентов и ресурсов. Воронцов К. В., Лексин В. А. Тезисы доклада на конференции ММРО-13.
- Технология персонализации на основе выявления тематических профилей пользователей и ресурсов Интернет. Лексин В. А. Диплом магистра.

2006

- Персонализация контента на основе оценок сходства пользователей и ресурсов сети интернет. Тезисы конференции МФТИ.
- Система имитационного моделирования транспортной сети аэропорта AirForS. Чехович Ю. В., Ефимов А. Н., Лексин В. А., Романов М. Ю., Рудева А. В., Громов С. А., Яминов Р. И. Научно-теоретический журнал «Искусственный интеллект» №.2'2006.
- Выявление и визуализация метрических структур на множествах пользователей и ресурсов Интернет. К. В. Воронцов, К. В. Рудаков, В. А. Лексин, А. Н. Ефимов. Научно-теоретический журнал «Искусственный интеллект» №.2'2006.

2005

- Методы выявления взаимосогласованных структур сходства в системах взаимодействующих объектов. Лексин В. А. Диплом бакалавра.

Публикации 2008-2009

2009

- Критерии ветвления в иерархическом вероятностном латентном семантическом анализе. В. А. Лексин. Сборник трудов МФТИ.
- **Symmetrization and overfitting in probabilistic latent semantic analysis. Pattern Recognition and Image Analysis. 2009. Volume 19. Number 4. December 2009. Pp. 565-574.**
- Двухступенчатые модели и проблема переобучения в латентном семантическом анализе. В. А. Лексин.

2008

- The overfitting in probabilistic latent semantic models. ROAI.
- Переобучение в вероятностных латентных семантических моделях. Лексин В. А., Воронцов К. В. Тезисы доклада на конференции РОАИ-2008.

Выступления на конференциях и семинарах

- *3 ноября 2009*. Иерархический вероятностный латентный семантический анализ. Семинар, ВЦ РАН.
- *14 мая 2009*. Двухступенчатые модели и проблема переобучения в латентном семантическом анализе. Семинар, ВЦ РАН.
- *17 сентября 2008*. The overfitting in probabilistic latent semantic models. ROAI, Н.Новгород, стендовый доклад.
- *20 мая 2008*. Сходство пользователей и ресурсов. Выявления скрытых тематических профилей. Семинар, ВЦ РАН.
- *24 ноября 2007*. Оценивание сходства пользователей и ресурсов путем выявления скрытых тематических профилей. Конференция МФТИ, г. Долгопрудный, Моск. обл.
- *2 октября 2007*. Анализ клиентских сред: выявление скрытых профилей и оценивание сходства клиентов и ресурсов. ММРО-13, г. Зеленогорск Ленинградской обл.
- *25 ноября 2006*. Интеллектуальный анализ данных о сходстве пользователей и ресурсов Интернет. Конференция МФТИ, г. Долгопрудный, Моск. обл.
- *7 июня 2006*. Анализ данных о поведении пользователей сети Интернет. ИОИ-2006, г. Алушта, Крым, Украина.
- *22 ноября 2005*. Выявление и визуализация метрических структур на множествах пользователей и ресурсов Интернет. ММРО-12, г. Звенигород, Моск. обл.

План дальнейших исследований. Экспериментальная работа

- Участие в создании проекта ***"Полигон алгоритмов коллаборативной фильтрации"***.
- Исследование и разработка функционалов качества алгоритмов коллаборативной фильтрации для последующего применения их в полигоне.
- Тестирование алгоритма на различных исходных данных. Сравнение алгоритмов.

План дальнейших исследований. Теоретическая работа

- Алгоритмы построения иерархических профилей и алгоритмы динамического обновления профилей при пополнении исходных данных.
- Дальнейшая разработка критериев достаточности наблюдений для определения компонент профилей.
- Дальнейшее теоретическое исследование сходимости и переобучения алгоритма.
- Исследование и оптимизация методов задания начального приближения профилей. Учет априорной информации.
- Обобщение алгоритма на случай дискретных рейтингов.
- Изучение кластерных свойств алгоритма.
- ***Написание статьи (текст в процессе подготовки)***.
- ***Защита диссертации***.