На правах рукописи

ЛЕКСИН ВАСИЛИЙ АЛЕКСЕЕВИЧ

ВЕРОЯТНОСТНЫЕ МОДЕЛИ В АНАЛИЗЕ КЛИЕНТСКИХ СРЕД

01.01.09 — Дискретная математика
и математическая кибернетика

Автореферат диссертации на соискание ученой степени
кандидата физико-математических наук

Москва, 2011

Работа выполнена на кафедре интеллектуальных систем Московского физико-технического института (государственного университета)

Научный руководитель:	доктор физико-математических наук **Воронцов Константин Вячеславович**
Официальные оппоненты:	доктор физико-математических наук **Сенько Олег Валентинович**
	кандидат технических наук **Игнатов Дмитрий Игоревич**
Ведущая организация:	Учреждение Российской академии наук Научно-исследовательский институт системных исследований РАН

Защита диссертации состоится « ____ » _____ 2011 г. в _____ на заседании диссертационного совета Д 002.017.02 в Учреждении Российской академии наук Вычислительный центр им. А. А. Дородницына РАН по адресу: 119333, г. Москва, ул. Вавилова, д. 40.

С диссертацией можно ознакомиться в библиотеке ВЦ РАН.

Автореферат разослан « ____ » _____ 2011 г.

Учёный секретарь диссертационного совета
Д 002.017.02, д.ф.-м.н., профессор В. В. Рязанов

Общая характеристика работы

Актуальность темы исследования. В работе вводится понятие клиентской среды, позволяющее с единых позиций подходить к анализу транзакционных данных, возникающих в различных прикладных областях. Клиентская среда описывается тремя множествами: множеством субъектов (клиентов, пользователей), множеством объектов (ресурсов, услуг, товаров, документов и т.д.) и множеством транзакций. В типичном случае транзакция представляет собой запись о взаимодействии некоторого субъекта с некоторым объектом. В качестве приложений можно рассматривать клиентские среды интернет-магазинов, поисковых машин, электронных библиотек, социальных сетей, торговых сетей, операторов связи, и т. д. Целью анализа клиентских сред (АКС) является построение информационных сервисов для выявления предпочтений клиентов, формирования персональных рекомендаций, поиска релевантных объектов или субъектов, выявления и визуализации скрытых закономерностей.

В работе исследуются методы вероятностного латентного семантического анализа (PLSA), основанные на байесовской вероятностной модели данных. Для идентификации параметров модели по выборке транзакций применяется итерационный EM-алгоритм, максимизирующий функционал правдоподобия. Методы PLSA позволяют получать сжатые семантические описания для каждого объекта и каждого субъекта в виде вектора вероятностей тем, называемого в работе *профилем* соответствующего объекта или субъекта.

Хотя данный подход активно применяется уже более 10 лет,

оценки скорости сходимости EM-алгоритма именно для PLSA до сих пор не были получены. Кроме того, оставались открытыми вопросы формирования начальных приближений и влияния разреженности профилей на качество решения и скорость сходимости EM-алгоритма. Получение ответов на эти вопросы является актуальной задачей.

Цель работы. Целью работы является получение оценок скорости сходимости EM-алгоритма для вероятностного латентного семантического анализа, а также разработка методов улучшения сходимости и повышения качества восстановления тематических профилей по транзакционным данным.

Научная новизна. Впервые получены оценки скорости сходимости EM-алгоритма в PLSA, установлены условия суперлинейной сходимости, выявлены основные параметры, воздействуя на которые можно улучшить сходимость EM-алгоритма. Разработаны новые эвристические методы, позволяющие улучшить качество восстановления профилей и скорость сходимости итерационного EM-алгоритма. Предложена симметризованная модель PLSA и метод оценивания её параметров на основе нового двухступенчатого EM-алгоритма. Предложен способ формирования начальных приближений для EM-алгоритма на основе быстрой кластеризации исходных данных, в то время как в литературе обычно предлагается брать случайные начальные приближения. Предложена методика поддержания разреженности тематических профилей в ходе итераций. Описана общая методология анализа клиентских сред, включающая операции предварительной обработки данных, предварительной класте-

ризации, восстановления профилей, вычисления функций сходства между объектами и субъектами, формирование рекомендаций, их ранжирование и визуализацию в виде карт сходства.

Методы исследований. Для построения байесовской вероятностной модели взаимодействия клиентов и объектов и оценки параметров модели с помощью принципа максимизации взвешенного правдоподобия (МВП) применяются методы теории вероятности и математической статистики. При выводе формул EM-алгоритма применяются методы минимизации функций с ограничениями типа равенств. Для оценки сходимости EM-алгоритма используются свойства линейных пространств и операторных норм. При разработке эвристических методов улучшения сходимости применяются методы математической статистики и комбинаторного анализа.

Хотя работа является теоретической, ход исследования в значительной степени определялся по результатам экспериментов на реальных и модельных задачах анализа клиентских сред.

Результаты, выносимые на защиту.

1. Симметризованная модель вероятностного латентного семантического анализа и метод оценивания её параметров на основе двухступенчатого EM-алгоритма.

2. Асимптотическая оценка скорости сходимости EM-алгоритма и условия его суперлинейной сходимости.

3. Способы улучшения сходимости EM-алгоритма и повышения качества восстановления профилей.

Практическая и теоретическая ценность. Вопрос о сходимости итерационных методов оценивания параметров моделей по выборкам данных является одной из ключевых проблем в математической теории распознавания и классификации. Настоящая работа даёт решение данной проблемы для задач восстановления тематических профилей, которые можно рассматривать как специальный класс задач кластеризации.

Предлагаемые методы улучшения сходимости EM-алгоритма и повышения качества восстановления профилей (формирование начального приближения, поддержание разреженности профилей, оптимизация параметров на модельных данных, учет априорной информации и т.д.) направлены на непосредственное практическое применение. В работе приводятся результаты экспериментов на реальных данных, демонстрирующие практическую применимость предложенных методов.

Аппробация работы. Результаты работы неоднократно докладывались на научных семинарах отдела Интеллектуальных систем ВЦ РАН и на конференциях:

- международная конференция «Распознавание образов и анализ изображений: новые информационные технологии» РОАИ-9, Нижний Новгород, 2008 г. [4];
- 50-я научная конференция МФТИ, 2007 г. [5];
- всероссийская конференция «Математические методы распознавания образов» ММРО-13, 2007 г. [6];
- 49-я научная конференция МФТИ, 2006 г. [7];
- международная конференция «Интеллектуализация обработки информации» ИОИ-6, 2006 г. [8];

- всероссийская конференция «Математические методы распознавания образов» ММРО-12, 2005 г.

Публикации. По теме диссертации опубликовано 9 работ, в том числе в изданиях из списка, рекомендованного ВАК РФ — одна [3], 7 в трудах конференций.

Структура и объем диссертационной работы. Диссертация изложена на 95 страницах машинописного текста и состоит из введения, 4 глав, заключения и списка литературы. Список литературы состоит из 41 наименования.

Краткое содержание работы по главам

ВВЕДЕНИЕ

Во введении дается обоснование актуальности работы, формулируется цель и задачи диссертационной работы, обосновывается научная новизна полученных результатов.

ГЛАВА 1. Задачи и методология анализа клиентских сред

1.1. Структура исходных данных. Пусть заданы множество $\mathcal{U} = \{1, \ldots, U\}$ номеров клиентов (субъектов, пользователей) и множество $\mathcal{R} = \{1, \ldots, R\}$ номеров объектов (ресурсов, товаров, предметов). Записи вида «клиент u взаимодействовал с объектом r» будем называть транзакциями. В зависимости от предметной области это могут быть покупки, визиты, просмотры и т. д. Пусть каждая транзакция описывается элементом

множества $\mathcal{Y}$. Протоколом транзакций называется последовательность записей $\mathcal{D} = \{(u_i, r_i, y_i) : i = 1, \ldots, N_D\} \subseteq \mathcal{U} \times \mathcal{R} \times \mathcal{Y}$, где N_D — длина протокола.

Из протокола транзакций можно получить агрегированные данные в виде матрицы кросс-табуляции размера $U \times R$: $F = \|f_{ur}\|$, $f_{ur} = \text{aggr}\{(u, r, y_i) \in \mathcal{D}\}$, где aggr — некоторая операция агрегирования. Например, f_{ur} — число использований объекта r клиентом u. Конкретный вид операции агрегирования зависит от характера множества $\mathcal{Y}$ и прикладной задачи.

1.2. Прикладные задачи, потребности. Перечисляются основные постановки задач в анализе клиентских сред: выдать оценку объекта r для клиента u; выдать клиенту u ранжированный список рекомендуемых объектов; построить по объекту r ранжированный список схожих с ним объектов; выявить тематику интересов клиента u; кластеризовать множество клиентов по интересам; визуализировать кластерную структуру клиентской среды. В результате их формализации возникают задачи прогнозирования незаполненных ячеек f_{ur}, оценивания функций сходства $K(u, u')$, $K(r, r')$, $K(u, r)$ между клиентами и объектами, формирования сжатых описаний клиентов и объектов в терминах латентных интересов или тем, одновременной кластеризации множеств клиентов и объектов.

В работе перечислены некоторые области применения технологии анализа клиентских сред: рекомендующие системы, анализ текстов, поисковые машины, интернет-магазины.

1.3. Методология анализа клиентских сред. В данном разделе рассматриваются отдельные методы анализа данных и

способы их совместного применения, в совокупности образующие методологию анализа клиентских сред. Описываются входные и выходные данные для каждого из методов.

ГЛАВА 2. Обзор методов коллаборативной фильтрации

В англоязычной литературе смежной с анализом клиентских сред областью исследований является коллаборативная фильтрация. В данной главе представлен обзор известных методов коллаборативной фильтрации. В разделе **2.1** описываются корреляционные методы: метод корреляции Пирсона, метод линейного сходства и точный тест Фишера. В разделе **2.2** рассматриваются методы, основанные на латентных моделях: латентный семантический анализ, вероятностный латентный семантический анализ и иерархический вероятностный латентный семантический анализ. В разделе **2.3** рассматриваются различные подходы к формированию начальных приближений в EM-алгоритме, более подробно описывается метод семплирования по небольшим подвыборкам объектов.

ГЛАВА 3. Вероятностный латентный семантический анализ

В главе рассматривается метод вероятностного латентного семантического анализа (PLSA), его модификации и свойства. Исследуются условия сходимости и предлагаются эвристические методы улучшения EM-алгоритма для PLSA.

3.1. Восстановление скрытых профилей клиентов и объектов. В разделе описываются две модификации стандартного вероятностного латентного семантического анализа.

Пусть $\mathcal{T} = \{1, \ldots, T\}$ — множество номеров тем объектов (интересов клиентов), T — число тем. Предполагается, что $T \ll R$ и $T \ll U$.

Рассмотрим вероятностное пространство на $\mathcal{U} \times \mathcal{R}$ с функцией распределения $p(u, r)$. В основе PLSA лежит вероятностная модель, которая может быть представлена в трех эквивалентных видах согласно формуле полной вероятности:

$$p(u,r) = \sum_{t \in \mathcal{T}} p_u p_{tu} q(r|t) = \tag{1}$$

$$= \sum_{t \in \mathcal{T}} q_r q_{tr} p(u|t) = \tag{2}$$

$$= \sum_{t \in \mathcal{T}} p_t q(r|t) p(u|t), \tag{3}$$

где $p_{tu} = p(t|u)$ — вероятность интереса клиента u к теме t; $q_{tr} = q(t|r)$ — вероятность того, что объект r относится к теме t; $p_u = p(u)$ — априорная вероятность клиента u; $q_r = q(r)$ — априорная вероятность объекта r; $p_t = p(t)$ — априорная вероятность темы t.

Согласно формуле Байеса, апостериорные вероятности распределения клиентов и объектов по темам имеют вид

$$p(u|t) = p_{tu} p_u / p_t, \quad q(r|t) = q_{tr} q_r / p_t.$$

Вектор $\mathbf{p}_u = (p_{tu} : t \in \mathcal{T})$ назовем профилем клиента $u \in \mathcal{U}$, а вектор $\mathbf{q}_r = (q_{tr} : t \in \mathcal{T})$ — профилем объекта $r \in \mathcal{R}$.

Задача вероятностного латентного семантического анализа — оценить по транзакционным данным $\mathcal{D}$ или по матрице кросс-табуляции F профили клиентов и объектов, используя вероятностную модель (1)-(3).

Априорные вероятности легко оценить по выборке данных: $p_u = \frac{S(u)}{S},\ q_r = \frac{S(r)}{S},\ S(u) = \sum\limits_{r\in\mathcal{R}} f_{ur},\ S(r) = \sum\limits_{u\in\mathcal{U}} f_{ur},\ S = \sum\limits_{r\in\mathcal{R}}\sum\limits_{u\in\mathcal{U}} f_{ur}$. Для нахождения неизвестных параметров $p_{tu},\ q_{tr}$ и p_t воспользуемся принципом максимума взвешенного правдоподобия:

$$l = \sum_{(u,r)\in\mathcal{U}\times\mathcal{R}} f_{ur} \ln p(u,r) \to \max_{\{p_{tu},q_{tr},p_t\}} \tag{4}$$

при ограничениях неотрицательности $p_{tu} \geqslant 0,\ q_{tr} \geqslant 0$ и нормировки $\sum\limits_{t\in\mathcal{T}} p_{tu} = 1,\ \sum\limits_{t\in\mathcal{T}} q_{tr} = 1,\ \sum\limits_{u\in\mathcal{U}} p_{tu}p_u = \sum\limits_{r\in\mathcal{R}} q_{tr}q_r = p_t$. Для решения данной оптимизационной задачи используется EM-алгоритм. Скрытыми переменными в EM-алгоритме являются апостериорные вероятности того, что клиент u, выбирая объект r, интересовался темой t:

$$h(t|u,r) = \frac{p_{tu}q_{tr}/p_t}{\sum\limits_{\tau\in\mathcal{T}} p_{\tau u}q_{\tau r}/p_\tau}. \tag{5}$$

Оптимизационная задача (4) решается аналитически, и её решение выражается через скрытые переменные:

$$\begin{aligned} p_{tu} &= \frac{1}{S(u)} \sum_{r\in\mathcal{R}} f_{ur} h(t|u,r), \\ q_{tr} &= \frac{1}{S(r)} \sum_{u\in\mathcal{U}} f_{ur} h(t|u,r), \\ p_t &= \frac{1}{S} \sum_{(u,r)\in\mathcal{U}\times\mathcal{R}} f_{ur} h(t|u,r). \end{aligned} \tag{6}$$

EM-алгоритм состоит из итерационного повторения двух шагов: E-шага (5) и M-шага (6). В качестве начального приближения задаются параметры p_{tu} и q_{tr}, параметры p_t вычисляются из условий нормировки. Итерации продолжаются, пока не произойдёт стабилизация значений параметров и/или значений правдоподобия.

Отличие предложенного метода от классического, описанного в работе Хофмана (2003), заключается в том, что здесь на каждом M-шаге непосредственно оцениваются компоненты профилей p_{tu} и q_{tr}, а не апостериорные вероятности $p(u|r)$ и $q(r|t)$, через которые профили должны ещё вычисляться по формуле Байеса.

Далее в работе предлагается симметризованный метод, основанный на вероятностных моделях (1) и (2) и представляющий собой двухступенчатый итерационный процесс, в котором профили p_{tu} и q_{tr} оцениваются поочередно, используя EM-алгоритм, подобный описанному выше. При оценке профилей p_{tu} профили q_{tr} считаются фиксированными, затем, наоборот, фиксируются значения p_{tu} и производится оценка профилей q_{tr}. Это позволяет задавать начальные приближения только для профилей клиентов, либо только для профилей объектов.

3.2. Оценка скорости сходимости EM-алгоритма. В данном разделе исследуются вопросы сходимости EM-алгоритма в методе PLSA.

Пусть $\mathbf{p}'_u$ и $\mathbf{q}'_r$ — профили после выполнения M-шага. Для EM-алгоритма в PLSA доказаны следующие теоремы:

Теорема 1. *На каждой итерации справедливы равенства:*

$$\begin{aligned} \mathbf{p}'_u - \mathbf{p}_u &= P_u \frac{\partial l}{\partial \mathbf{p}_u},\ u \in \mathcal{U}, \\ \mathbf{q}'_r - \mathbf{q}_r &= Q_r \frac{\partial l}{\partial \mathbf{q}_r},\ r \in \mathcal{R}, \end{aligned} \tag{7}$$

где $P_u = \frac{1}{S(u)}\big(\operatorname{diag}(p_{1u}, \ldots, p_{Tu}) - \mathbf{p}_u \mathbf{p}_u^\intercal\big)$;

$Q_r = \frac{1}{S(r)}\big(\operatorname{diag}(q_{1r}, \ldots, q_{Tr}) - \mathbf{q}_r \mathbf{q}_r^\intercal\big)$.

Теорема 2. *Матрицы P_u и Q_r положительно полуопределены для всех $u \in \mathcal{U}$ и $r \in \mathcal{R}$.*

Обозначим вектор всех параметров модели на k-ой итерации через $\theta = [\mathbf{p}_1^\intercal, \ldots, \mathbf{p}_U^\intercal, \mathbf{q}_1^\intercal, \ldots, \mathbf{q}_R^\intercal]^\intercal$ и рассмотрим блочно-диагональную матрицу $P = \operatorname{diag}\{P_1, \ldots, P_U, Q_1, \ldots, Q_R\}$. Тогда выражения (7) примут вид

$$\theta' - \theta = P \frac{\partial l}{\partial \theta},$$

где θ' — значения вектора параметров на следующем шаге EM-алгоритма.

Из утверждения теоремы 2 следует, что матрица P положительно полуопределена, что имеет следующую геометрическую интерпретацию: разность векторов $\theta' - \theta$ на каждом шаге EM-алгоритма имеет неотрицательную проекцию на градиент правдоподобия. Это показывает тесную связь EM-алгоритма с градиентным методом, когда на каждом шаге движение идет в направлении градиента с некоторым выбранным шагом. Известно, что для градиентного метода гарантирована сходимость к локальному максимуму правдоподобия.

В следующей теореме утверждается, что EM-алгоритм имеет асимптотически линейную скорость сходимости.

Теорема 3. *Пусть θ^* — локальный максимум $l(\theta)$, $\theta \to \theta^*$ при $k \to \infty$, где k — номер итерации; в некоторой окрестности θ^* Гессиан $H(\theta) = \frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^\intercal}$ существует и отрицательно определен. Тогда справедлива оценка*

$$\lim_{k\to\infty} \frac{\|\theta' - \theta^*\|}{\|\theta - \theta^*\|} \leqslant r_c,$$

где $r_c = \|E^\intercal\big(I + P(\theta^*)H(\theta^*)\big)\| \leqslant \sqrt{1 + \lambda_M^2 - 2\lambda_m}$, λ_M и λ_m — минимальное и максимальное собственные значения положительно полуопределенной матрицы $M = -E^\intercal P(\theta^*)H(\theta^*)E$, E — произвольный ортонормированный базис линейного подпространства

$$\Theta = \Bigg\{\hat\theta = \theta' - \theta : \sum_{t\in\mathcal{T}} \hat p_{tu} = 0,\ u \in \mathcal{U};\ \sum_{t\in\mathcal{T}} \hat q_{tr} = 0,\ r \in \mathcal{R};$$
$$\sum_{u\in\mathcal{U}} p_u \hat p_{tu} = \sum_{r\in\mathcal{R}} q_r \hat q_{tr},\ t \in \mathcal{T}\Bigg\};$$

Скорость сходимости критически зависит от степени обусловленности матрицы PH. Если матрица обусловлена плохо, то сходимость алгоритма не гарантируется.

В следующей теореме утверждается, что если все векторы $\big(h(t|u,r)\colon t \in \mathcal{T}\big)$, для которых $f_{ur} \neq 0$, содержат строго по одному ненулевому элементу, то алгоритм имеет суперлинейную скорость сходимости.

Теорема 4. *Пусть $h(t|u,r) \in \{0,1\}$ для всех $t \in \mathcal{T}$, $u \in \mathcal{U}$ и $r \in \mathcal{R}$ в точке θ^*. Тогда $r_c = 0$.*

Для полученного теоретического результата можно провести следующую содержательную интерпретацию: если события

(u, r) выбора объекта r клиентом u всегда обусловлены интересом клиента u к одной и той же теме $t \in \mathcal{T}$, то EM-алгоритм сходится с суперлинейной скоростью.

3.3. Методы улучшения сходимости EM-алгоритма. В разделе рассматриваются два эвристических метода улучшения сходимости EM-алгоритма.

1. Метод задания начального приближения профилей по протоколу транзакций, состоящий из следующих шагов.

- Выделяется небольшая подвыборка объектов (метод семплирования). Подвыборка может быть либо случайной, либо включать объекты, для которых априори известно, что они относятся к разным темам.
- Из полученных на предыдущем шаге объектов отбираются T представительных объектов. Каждый представительный объект отвечает за соответствующий компонент профиля, что позволяет произвести содержательную интерпретацию компонент профилей.
- Начальные приближения компонент профилей объектов оцениваются пропорционально сходству выбранного объекта с соответствующим представительным объектом, оцененного методом сглаженной корреляции.

2. Метод увеличения разреженности профилей и векторов скрытых вероятностей путем обнуления близких к нулю компонент. На каждой EM-итерации отбрасываются (обнуляются) близкие к нулю скрытые переменные $h(t|u, r)$ и компоненты профилей по заданным порогам. Данный подход позволяет существенно сократить объем хранимых в памяти данных.

ГЛАВА 4. Эксперименты

В данной главе описываются вычислительные эксперименты на модельных и на реальных данных. Для оценки работы алгоритма и оптимизации параметров вводятся специальные функционалы качества. На модельных данных функционал качества определяется как среднеквадратичное отклонение (СКО) компонент профилей, полученных на выходе EM-алгоритма, от компонент модельных профилей, используемых для генерации протокола. На реальных данных функционал качества определяется как число ошибок классификации частично размеченной выборки объектов методом k ближайших соседей (kNN). В экспериментах показывается, что предложенные эвристические методы улучшают качество EM-алгоритма и повышают скорость сходимости.

В разделе **4.1** описываются используемые наборы данных. В разделе **4.2** исследуются свойства и способы улучшения сходимости EM-алгоритма на модельных данных. Для исследования влияния начального приближения профилей на результат EM-алгоритма берется идеальное начальное приближение (модельные профили), с наложенным на него аддитивным шумом заданной амплитуды σ. Выяснилось, что скорость и качество сходимости EM-алгоритма критически зависят от данного параметра. Для проверки условия суперлинейной сходимости был проведен эксперимент, в котором удалось выяснить, что алгоритм сходится за 4 шага при выполнении условия теоремы и фиксированных параметрах моделирования протокола. Задание начального приближения профилей по протоколу увеличивает скорость сходимости в 3 раза и уменьшает СКО от модельных профилей на 40%. Метод обнуления малых компонент

скрытых переменных и профилей увеличивает скорость сходимости в 2 раза и улучшает качество на 15%.

В разделе **4.3** описаны эксперименты на реальных данных. На данных поисковой машины «Яндекс» показывается, что симметризованный вероятностный латентный семантический анализ увеличивает скорость сходимости в 2 раза и улучшает качество по kNN на 25%. Далее на данных поисковой машины строится полная и локальная карты сходства объектов. На данных о действиях клиентов Интернет-магазина исследуется метод задания начального приближения профилей клиентов и ресурсов по протоколу. В результате качество улучшилось в 3 раза и СКО от модельных профилей уменьшилось на 30% по сравнению со случайным начальным приближением.

В **заключении** приведены основные результаты работы.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

[1] Лексин В. А. Методы улучшения сходимости EM-алгоритма в вероятностном латентном семантическом анализе // Математические методы распознавания образов: 15-я Всерос. конф. Тезисы докл.— 2011. — С. 262–265.

[2] Лексин В. А. Критерии ветвления в иерархическом вероятностном латентном семантическом анализе // Труды 52-й научной конференции МФТИ, Т. 7, ч. 2. — 2009. — С. 76–79.

[3] **Leksin V. A. Symmetrization and overfitting in probabilistic latent semantic analysis // Pattern Recognition and Image Analysis. — Vol. 19, no. 4 — 2009. — Pp. 565–574.**

[4] Leksin V. A., Vorontsov K. V. The overfitting in probabilistic latent semantic models.// Pattern Recognition and Image Analysis: new information technologies (PRIA-9). Vol. 1. Nizhni Novgorod, Russian Federation. — 2008. — Pp. 393–396.

[5] Лексин В. А. Оценивание сходства пользователей и ресурсов путем выявления скрытых тематических профилей // Труды 50-й научной конф. МФТИ. Часть VII. Том 2. — 2007. — С. 104–106.

[6] Воронцов К. В., Лексин В. А. Анализ клиентских сред: выявление скрытых профилей и оценивание сходства клиентов и ресурсов // Математические методы распознавания образов: 13-я Всерос. конф. Тезисы докл. — 2007. — С. 488–491.

[7] Лексин В. А., Воронцов К. В. Персонализация контента на основе оценок сходства пользователей и ресурсов сети интернет // Труды 49-й научной конф. МФТИ. Часть VII. Том 2. — 2006. — С. 276–277.

[8] Воронцов К. В., Рудаков К. В., Лексин В. А., Ефимов А. Н. Выявление и визуализация метрических структур на множествах пользователей и ресурсов Интернет // Научно-теоретический журнал «Искусственный интеллект». №2. — 2006. — С. 285–288.

[9] Воронцов К. В., Рудаков К. В., Лексин В. А., Ефимов А. Н. О метрических структурах на множествах пользователей и ресурсов Интернет // Интеллектуализация обработки информации: междунар. конф. Тезисы докл. — 2006. — С. 46–48.

В работах с соавторами лично соискателем сделано следующее:

- разработаны и реализованы алгоритмы обработки логов поисковой машины «Яндекс» [8, 9];
- проведены эксперименты по оцениванию метрики на множестве объектов по точному тесту Фишера [7, 8];

- разработаны методы оценивания качества метрик, проведены эксперименты по восстановлению скрытых профилей и построению карт сходства объектов [6, 4].