САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

На правах рукописи

КИСЕЛЁВА Юлия Евгеньевна

МЕТОДЫ ГРУППИРОВКИ И СТРУКТУРИЗАЦИИ ПОИСКОВЫХ ЗАПРОСОВ И ИХ РЕАЛИЗАЦИЯ

05.13.11 — математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
Диссертации на соискание ученой степени
кандидата физико-математических наук

Санкт-Петербург
2011

Работа выполнена на кафедре информатики математико-механического факультета Санкт-Петербургского государственного университета.

Научный руководитель: доктор физико-математических наук,
профессор НОВИКОВ Борис Асенович

Официальные оппоненты: доктор физико-математических наук,
профессор ШЕВЛЯКОВ Георгий Леонидович
(Санкт-Петербургский государственный
политехнический университет)

кандидат технических наук,
БРАСЛАВСКИЙ Павел Исаакович
(Уральский государственный университет)

Ведущая организация: Южно-Уральский государственный университет

Защита диссертации состоится "___"_____ 2011 года в ____ часов на заседании совета Д212.232.51 по защите докторских и кандидатских диссертаций при Санкт-Петербургском государственном университете по адресу: 198504, Санкт-Петербург, Старый Петергоф, Университетский пр., д. 28, математико-механический факультет Санкт-Петербургского государственного университета, ауд. 405.

С диссертацией можно ознакомиться в Научной библиотеке им. М.Горького Санкт-Петербургского государственного университета по адресу: 199034, Санкт-Петербург, Университетская наб., д. 7/9.

Автореферат разослан "____" _____ 2011 года.

Ученый секретарь
диссертационного совета
доктор физико-математических наук,
профессор Даугавет И.К.

Общая характеристика работы

Актуальность темы. Исследованиям в области анализа поисковых запросов уделяется много внимания в последние годы. Этому способствуют многие факторы, среди которых:

- общедоступность интернета для пользователей;
- увеличение объема полезной для пользователей информации в интернет-пространстве.

Данные факторы приводят к тому, что пользователи все чаще прибегают к поиску нужной им информации в интернете, и свои потребности они формулируют в виде запросов с *«ключевыми словами» (keyword queries),* и, как следствие, объем обрабатываемых поисковых запросов значительно увеличивается каждый год. В результате накапливаются большие по объему журналы, содержащие поисковые запросы пользователей *(search query logs).* Однако, любые коллекции данных бесполезны, если не существует методики для их анализа.

Запросы пользователей важнейшая для владельцев интернет-ресурсов информация. Так как выводы, полученные путем анализа поисковых запросов, потенциально могут улучшить качество поиска, так как они помогают лучше понять интересы пользователей. И с учетом полученных знаний *поисковые машины (search engine)* будут показывать наиболее релевантные пользователю документы.

Одной из основных проблем анализа поисковых запросов является *неоднозначность (ambiguity)* используемых в них слов. Один из классических примеров подобной неоднозначности является запрос *“jaguar”*. В данном случае непонятно, о чем конкретно искали информацию: об автомобилях или о животных. Если же мы обладаем знаниями об интересах пользователя, который ввел неоднозначный запрос, мы легко сможем определить, какого рода информацию он хотел узнать.

Также большое внимание уделяется методам, которые позволяют преобразовывать неструктурированный запрос пользователя с *«ключевыми словами» (keyword queries)* в структурированный. Основная причина популярности подобных методов заключается в том, что большая часть интернет-данных изначально содержатся в структурированных базах данных. И знание структуры запроса значительно облегчает поиск релевантных ответов.

Для обучения модели анализа запросов, которая получает из запроса структуру, необходимо составить обучающее множество, в котором каждый запрос описывается *векторами признаков (feature vector) или просто признаками* – наборами числовых параметров, отражающих свойства характеристик

запроса. Вектора признаков принимают значения в пространстве признаков. Задав метрику в подобном пространстве, можно сравнивать запросы друг с другом, вычисляя расстояние между соответствующими им векторами. Методики для создания обучающего множества и построения векторов признаков являются ядром любой системы анализа запросов. Качество системы анализа поисковых запросов в основном зависит от выбора обучающего множества и признаков, а также метрик для их сравнения.

Традиционным подходом для создания системы анализа запросов является *обучение «с учителем» (supervised learning)*, но данный метод представляется достаточно трудоемким и дорогостоящим, так как требует обучающего множества, составленного вручную.

В последние годы так же были разработаны методы, использующие *«частичное обучение с учителем» (semi-supervised learning)*, которые используют небольшое по размеру обучающее множество на первом этапе обучения, а затем итеративно добавляют наиболее хорошие предсказания, таким образом, расширяется обучающее множество.

В настоящее время существуют огромные объемы данных, которые содержат журналы щелчков пользователей. И естественно предположить, что на основе этих данных есть возможность создать обучающее множество автоматически, без использования работы асессоров, составляющих обучающее множество вручную. Данная концепция получила название обучение «без учителя» *(unsupervised learning)*.

Цели диссертационной работы. Основной целью работы является создание высокоэффективных, с точки зрения качества результата, методов обучения «без учителя» для построения систем анализа поисковых запросов. Для достижения поставленной цели были выделены следующие задачи:

- Разработка эффективной метрики, базирующейся на журналах запросов, которая служит инструментом для нахождения групп пользователей, характеризующихся похожими интересами.
- Разработка метода для автоматического построения обучающего множества, данная задача обуславливается желанием не использовать дорогостоящие и трудоемкие методы составления обучающего множества вручную. Данный метод в качестве входных данных использует *журналы щелчков пользователей (user clicks logs)* и *базу данных с описанием продуктов (product data base)*, которые ищут и на которые щелкают пользователи.
- Построение эффективных *признаков* для обучения вероятностной модели сегментации запросов, которая преобразует неструктурированный запрос в структурированный.

Основные результаты. В работе получены следующие основные результаты:

1. Метод для группировки пользователей интернета на основе их запросов, основанный на построении метрики для нахождения пользователей со схожими интересами.
2. Реализованный прототип системы для группировки пользователей по интересам. Эксперименты для оценки достоверности полученных метрик проведены на реальных англоязычных запросах пользователей.
3. Новый метод для автоматического составления обучающего множества, состоящего из поисковых запросов коммерческого типа, на основе сопоставления журналов щелчков пользователей и базы данных продуктов.
4. Метод для сегментации поисковых запросов коммерческого типа, обученный на автоматически полученном тренировочном множестве, с возможностью регулировать степень доверия каждого предсказания. На основе этого метода построены вероятностные модели для сегментации запросов.
5. Реализованный прототип системы для сегментации запросов, работающий на основе категории введенной пользователем и вероятностной модели, построенной в результате обучения.
6. Проведены эксперименты с системой сегментации запросов на реальных данных и получены высокие экспериментальные оценки полноты, точности предложенного метода.

Научная новизна. Научной новизной обладают следующие результаты работы:

- предложенный метод для построения метрики, используемой для нахождения похожих пользователей путем анализа журналов их поисковых запросов;
- предложенный метод для автоматического построения обучающего множества, которое используется в процессе обучения вероятностной модели для сегментации запросов;
- предложенный метод построения эффективных признаков для модели обучения сегментации запросов, которая преобразует неструктурированный запрос в структурированный запрос;
- разработанная система для сегментации запросов, которая обучается «без учителя».

Теоретическая ценность и практическая значимость. Главными причинами внедрения методов анализа поисковых запросов являются:

- улучшение качества поиска;
- улучшение ранжирования результатов поиска;
- структуризация запросов с «ключевыми словами»;
- персонализация информации.

Предложенные в работе методы также находят широкое применение в различных областях интернет-индустрии, таких как:

- *вертикальный поиск.* Системы вертикального поиска ориентированы на конкретную область и позволяют осуществлять глубокий поиск именно по данной тематике. Информация об интересах пользователей и знание структуры запроса помогают улучшить поиск.
- *интернет-магазины*, для которых знания об интересах пользователей представляются жизненно важными, так как они стремятся показать пользователю наиболее релевантный продукт.
- *рекламные интернет компании*, для которых знания об интересах пользователей также являются необходимыми, так как они стремятся показать рекламу, соответствующую интересам пользователей, и таким образом избавить пользователей от ненужной и нерелевантной для них информации.

В рамках данной работы был реализован прототип системы сегментации запросов, его эффективность работы была доказана методом экспертной оценки. Этот прототип используется компанией *Shopping.com (Ebay.com)* в качестве инструмента преобразования неструктурированных запросов в структурированные. Также практическую ценность имеет предложенный метод для группировки схожих по интересам веб-пользователей, который основан на журналах их поисковых запросов.

Апробация работы. Основные результаты диссертации докладывались на следующих конференциях и семинарах:

- на десятой Всероссийской научной конференции *«Электронные Библиотеки: перспективные методы и технологии, электронные коллекции»* RCDL 2008, на которой работа была награждена, как лучший студенческий постер;
- на PhD Workshop двенадцатой *Восточно-Европейской Конференции по Базам Данных и Информационным системам* ABDIS 2008;
- на *Workshop Distributed Intelligent Systems and Technologies proceedings* 2009*;*

- на третьей и четвертой конференциях *Молодых Ученых при Российской Школе по Информационному Поиску* RUSSIR 2009 и RUSSIR 2010, на последней работа была награждена, как лучшая статья;
- на двадцатой Международной Конференции *World Wide Web* WWW 2010;
- на семинарах группы исследования методов организации информации при лаборатории исследования операций НИММ.

Публикации. Основные результаты диссертации были опубликованы в работах: 1-6. Статья 1 опубликована в журнале, входящем в список ВАК. Статьи 3 и 6 написаны в соавторстве. В статье 3 соискателю принадлежит идея метода автоматического построения обучающего множества и метод для расширения пространства признаков при построении вероятностной модели для сегментации запросов. В статье 6 соискателю принадлежит идея использования средней и медианной статистики.

Структура и объем диссертации. Диссертация состоит из введения, 3 глав, заключения и списка литературы. Общий объем диссертации составляет 99 страниц машинописного текста. Библиография содержит 73 наименования. Рисунки и таблицы нумеруются последовательно.

Содержание работы

Во **введении** содержится предварительная информация о предмете исследования, обосновывается актуальность тематики диссертационной работы и кратко излагаются ее основные результаты.

В **первой главе** «Методы анализа поисковых запросов пользователей» представлены направления и задачи области исследования, рассматриваются основные алгоритмы анализа поисковых запросов, обсуждаются известные на сегодняшний день подходы и методы оценки, заимствованные из информационного поиска.

В разделе 1.1 представлены традиционные модели информационного поиска: векторные модели и метод для вычисления весов слов. Данные модели адаптированы для решения задач поставленных в диссертации.

В разделах 1.2 – 1.4 описаны методы оценки, применяемые в задачах информационного поиска, и тестовые наборы данных. Представленные методики были применены и для оценки разработанных в диссертации методов анализа поисковых сессий пользователей. В том числе в разделе 1.2.1 представ-

лено описание инструмента *Amazon Mechanical Turk*[1], который был использован при составлении тестового множества для экспертной оценки, которое использовалось в дальнейшем для определения качества экспериментов.
В разделе 1.5 обсуждается применимость методов группировки интернет-пользователей по интересам в таких областях исследований как:

- персонализация информации,
- поиск шаблонов.

Также в данном разделе представлены методы группировки пользователей.
В разделе 1.6 рассматриваются некоторые из наиболее известных на сегодняшний день вероятностных моделей на графах, которые используются для разметки последовательностей слов. Обсуждается их применимость в задаче анализа поисковых запросов.
В разделе 1.7 содержится описание системы *WordNet*, которая была использована в диссертации для фильтрации опечаток в запросах.
Во **второй главе** «Группировка пользователей по интересам» предлагается новый метод построения метрики для составления групп пользователей по интересам. Данная методика в качестве входных данных использует журналы, в которых зафиксированы поисковые сессии пользователей. Также в главе представляется доказанная экспериментально эффективность предложенных методов.
Журналы поисковых сессий пользователей, в большинстве своем (80,6%) состоят из информационных запросов, которые отражают заинтересованность в конкретной информации. И это предположение позволяет полагать, что поисковые запросы отражают интересы пользователей.
Раздел 2.1 описывает детальную классификацию поисковых запросов с выкладками о статистике (процентное соотношение каждого класса запросов).
В разделе 2.2 представлены две метрики, позволяющие измерять схожесть между журналами запросов:

- *усредненная мера близости (УМБ);*
- *максимизированная мера близости (ММБ).*

В разделе 2.2.1 описывается метод построения УМБ, которая порождает пространство слов, состоящее из объединения всех запросов:

$$R^N : \{t_i\} ,$$

где N – это количество уникальных элементов и $\{t_i\}$ - это множество всех уникальных слов, которые присутствуют в журналах запросов, и именно они

[1] https://www.mturk.com/mturk/welcome

являются координатами обозначенного векторного пространства. Вектор запросов пользователей представлен в следующем виде:

$$d_j = (w(t_1), w(t_2), \ldots, w(t_j), \ldots, w(t_N)),$$

где $w(t_j)$ - это вес – *tf*idf (term frequency and inverse document frequency).* Для определения расстояния между пользователями используется косинусная мера близости, которая высчитывается по следующей формуле:

$$M(d_i, d_j) = \cos(d_i, d_j).$$

В разделе 2.2.2 описывается метод для построения ММБ, для которой каждый пользователь должен быть представлен в виде набора векторов для его запросов. Каждый вектор запросов *j*-ого пользователя имеют следующий вид:

$$Q_{jm} = (w(t_1), w(t_2), \ldots, w(t_l), \ldots, w(t_N)),$$

где $w(t_l)$ - это вес слова l в запросе Q_{im}. И документ *j*-ого пользователя в векторном пространстве характеризуется набором его запросов и представляется в виде:

$$d_j = \{Q_{j1}, Q_{j2}, \ldots, Q_{jn}\}.$$

И мера близости между двумя пользователями определяется по следующему принципу:

$$M(d_i, d_j) = max\{cos(Q_{jk}, Q_{im})\}.$$

В разделе 2.3 описывается набор данных для эксперимента. Раздел 2.4 описывает методику «очистки» запросов с помощью системы *WordNet.*

Раздел 2.5 содержит описание и анализ результатов эксперимента по оценке эффективности построенных метрик, которые представлены в таблице 3 в диссертации.

Согласно полученным результатам метрика УМБ показала высокие значения точности (83%), в то время как для ММБ точность не столь высока (77%). Однако значения полноты для ММБ в два раза выше (60%), чем для УМБ (30%). Поэтому метрика ММБ является наиболее приемлемой для задачи построения групп пользователей со схожими интересами.

В **третьей главе** «Сегментация поисковых запросов» обсуждается методика машинного обучения «без учителя» для построения вероятностной модели,

на основе автоматически полученного обучающего множества. Данная модель может конвертировать запросы с «ключевыми словами» в структурированные запросы.

В разделе 3.1 содержится описание практических проблем, которые сводятся к задаче сегментации. Обсуждается задача сегментации поисковых запросов пользователей, которая решается в данной работе. Она визуализируется на примере запроса *"sony vaio 12 200gb",* который в результате работы метода преобразуется в следующую структуру:

$$\text{sony} \rightarrow \textit{бренд},\ \text{vaio} \rightarrow \textit{модель},\ 14 \rightarrow \textit{размер дисплея},$$
$$200\text{gb} \rightarrow \textit{объем жесткого диска.}$$

В разделе 3.2 приводится обзор существующих методов сегментации запросов и подробно представлены следующие методики:

- *Условных случайных полей (УСП)(Conditional Random Fields)*;
- *Скрытой Марковской Модели (СММ) (Hidden Markov Model).*

Метод УСП позволяет моделировать зависимость между *полями* (в нашем случае это атрибуты) и *наблюдениями* (в нашем случае слова из запроса). Также в разделе обсуждаются известные методики обучения «с учителем» и «частичное обучение с учителем». Обосновывается применение внешних источников данных для автоматического составления обучающего множества, эффективного которого была доказана в предыдущих работах.

В данной работе в качестве внешних источников данных использовались:

- журналы щелков пользователей *(user click logs);*
- база данных продуктов *(products data bases).*

В разделе 3.3 сформулированы следующие требования к разработанной системе сегментации запросов:

- Обучающее множество строится на основе автоматически составленного словаря атрибутов, который получен на основе информации из журналов щелчков пользователей и базы данных.
- Автоматически полученное обучающее множество расширено с помощью метода генерации *«синтетических запросов».* Данное нововведение помогает нам «бороться» с разреженными данными.
- Признаки для УСП расширены, по сравнению описаниями в других работах, путем добавления *«словарных признаков слов»*, которые мы получаем автоматически путем анализа базы данных. Данные признаки необходимы для улучшения качества предсказаний УСП.

Раздел 3.4 содержит описание процесса автоматического маркирования запросов. Данная методика последовательно описана в разделах 3.4.1-3.4.4. В

разделе 3.4.1 подробно описываются входные данные для метода, а именно журналы поисковых сессий и база данных продуктов. Обсуждается метод для составления промаркированного множества, на основе косинусной меры близости. В качестве функции присвоения весов для слов, используется модифицированный *tf*idf (term frequency and inverse document frequency).* Модификация состоит в том, что *«документ»* в нашем случае - это комбинация всех слов, описывающих данных атрибут, среди всех продуктов из базы данных, на которые щелкали пользователи. Таким образом, мы получаем *тематические документы*. Рассмотрим на примере «*документа брендов»,* который содержит все слова и их частоты, которые встречались в описании атрибута *«бренд».* Он будет выглядеть следующим образом:

< ‘dell’:140, ‘lenovo’:90, ‘asus’:70>.

Если же слово не определено ни одним атрибутом, то ему присваивается значение *«неизвестный»*.

Список атрибутов, полученный путем объединения всех атрибутов из автоматически промаркированных запросов, будем называть *целевым.*

В разделе 3.4.2 описывается словарь брендов и их аббревиатур, который используется в «улучшенном» процессе маркирования запросов.

В разделе 3.4.3 представляется процесс уменьшения разреженности в обучающем множестве. И раздел 3.4.4 содержит описания критерия составления обучающего множества. А именно, в обучающее множество отбираются запросы, которые удовлетворят следующим двум свойствам:

- все пары слово-атрибут в запросе должны иметь присвоенное им значение косинусной меры не менее 0.3;
- запросы могут содержать слова, определенное атрибутом *«неизвестный»*, но, по крайней мере, два слова в запросе должны быть представлены *целевые атрибутами*.

В разделе 3.4.5 рассматривается новый метод построения *«синтетических запросов».* Построение «*синтетических запросов»* обуславливается необходимостью расширить обучающее множество.

В ходе эксперимента было построено распределение вероятностей перехода от одного атрибута к другому, необходимое для создания *«синтетических запросов»*. Данное распределение создано с помощью автоматически полученного обучающего множества, представленного в разделе 3.4. При построении используется алгоритм *«сглаживания».* То есть нулевые вероятности заменяются маленькими значениями для каждого атрибута, как показано в формуле (*) для атрибутов a_i. Данный метод применяется с целью создать

разнообразные *«синтетические запросы»*, путем увеличения вероятности перехода от одного атрибута к другому.
Описываемое распределение перехода строится с использованием двух специальных символов *«начало запроса»* и *«конец запроса»*. Согласно описанной методике, распределение для специального атрибута *«начало запроса»* выглядит следующим образом:

$$dist_{начало} = ((a_i = бренд,\ p_1 = 0.7),(a_2 = семейство,\ p_2 = 0.2),$$
$$(a_3 = объем\ памяти,\ p_3 = 0.1),\ldots(a_i = i\_атрибут,\ p_i = 0.01),\ldots), \quad (*)$$

При построении *«синтетических запросов»*, продукты, на которые щелкнули пользователи, рассматриваются один за другим. Каждый продукт P_i представлен в следующей форме:

$$P_i = \{название\ атрибута\ (a_j),\ описание\ этого\ атрибута\ (desc_j)\}.$$

«Синтетические запросы» создаются на основе описаний атрибутов и описанного выше распределения вероятностей перехода.
В разделе 3.6 представлена реализация метода для автоматического составления обучающего множества.
В разделе 3.6.1 показано подробное описание разработанной системы, которая состоит из трех основных частей:

1. *Автоматическое маркирование запросов.*
2. *Генерация обучающего множества.*
3. *Обучение модели для сегментации запросов.*

Раздел 3.6.1 описывает нормализацию базы данных продуктов путем выявления синонимов среди целевых атрибутов. Для этого процесса используется метод, описанный во второй главе.
Также в разделе 3.6.2 представлен метод для нормализации запросов.
Раздел 3.7 посвящен описанию процесса обучения модели для сегментации запросов. В разделе 3.7.1 обосновывается выбор вероятностной модели *Условные Случайные Поля (УСП)* для решения задачи сегментации.
Линейная цепь *УСП* получила широкое применение в задачах маркирования последовательностей, таких, как определение частей речи или извлечение информации.
В описываемой задаче имеется запрос, состоящий из n слов - $x = (x_1, x_2,\ldots,x_n)$ и $y = (y_1, y_2,\ldots,y_n)$ - это соответствующая ему последовательность атрибутов.

Условная вероятность между словом и атрибутом определена следующим образом:

$$p(y \mid x, \lambda) = \frac{1}{Z_\lambda(x)} exp(\sum_{i=1}^{n} \sum_j \lambda_j f_j(y_{i-1}, y_i, x, i)),$$

где функция $Z^{\lambda}(x)$ – это нормализующий фактор, j – это коллекция признаков и λ - это соответствующий вес.

В разделе 3.7.2 приведено подробное описание целевых атрибутов для системы сегментирования запросов.

Раздел 3.7.3 описывает набор признаков, используемых для обучения УСП. Данный набор признаков подразделяется на две основные группы:

- общие признаки;
- признаки, основанные на словаре. Они используют информацию из словарей, которые были изначально созданы для уменьшения разреженности обучающего множества. Данные словари были подробно описаны в разделе 3.4.

Обученная модель УСП[2] для сегментации предсказывает атрибут для слова с определенным *уровнем доверия.*

Уровень доверия - это коэффициент, значение которого принадлежит отрезку [0,1]. Он отражает *«уверенность»* метода в данном предсказании.

В разделе 3.8 приводятся описание следующих составляющих для проведения эксперимента:

- метрики качества;
- обучающее множество;
- тестовое множество.

В разделе 3.9 приводятся анализ результатов эксперимента. Процесс анализа был разделен на два этапа:

1. В разделе 3.9.1 проводится оценка автоматического маркирования запросов. Предложенный метод показал высокие оценки, а именно точность достигает 93 % и полнота 60%. В то время как точность и полнота для базы - 83% и 37% соответственно. Данный факт говорит о превосходстве представленного в диссертации метода.
2. В разделе 3.9.2 представляется описание результирующих оценок для метода сегментации запросов. Полученный метод показал высокие оценки: точность составляет 75 % и полнота - 53%. В то время как точ-

[2] Для обучения модели УПС используется имплементация, представленная в библиотеке Mallet (http://mallet.cs.umass.edu/)

ность и полнота для базы – 70% и 51 % соответственно. Приведенные измерения показывают превосходство разработанного метода.

Раздел 3.10 содержит основные выводы по исследованию методов сегментации запросов.

В **заключении** формулируется список основных результатов, полученных в диссертационной работе.

Публикации автора по теме диссертации

Статьи в журналах, рекомендованных ВАК:

1. Киселёва Ю. Е. Автоматическое сегментирование запросов интернет-магазинов. *Программные продукты и системы*. – 2010. – Вып. № 3 (91). – С. 129 – 131.

Другие публикации:

2. Киселёва Ю. Группировка пользователей интернета, основанная на истории их веб-сессий. *Труды 10-ой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции " RCDL'2008*. – 2008. - С. 405-407.
3. Julia Kiseleva, Eugene Agichtein, Qi Guo, Daniel Billsus, Wei Chai. Unsupervised Query Segmentation Using Click Data: Preliminary Result. In *processing of 19th International Word Wide Web Conference (WWW'2010)*. – 2010. - Pp. 1131-1132.
4. Julia Kiseleva. Grouping Web Users based on Query Log. In *processing of* 12th *East European Conference Advances in Databases and Information Systems (ADBIS'2008).* – 2008. - Pp. 184-190.
5. Julia Kiseleva. Unsupervised Query Segmentation Using Click Data and Dictionaries Information. *IV Российская летняя школа по информационному поиску RuSSIR'2010. Труды Четвертой Российской конференции молодых ученых по информационному поиску.* — 2010. — С. 6-13.
6. Mikhail Kalinkin, Julia Kiseleva, Nikolay Vyahhi, Bernhard Lang. Comparison of Machine Learning Techniques for Document Ranking Problem. In *processing of Workshop Distributed Intelligent Systems and Technologies proceedings.* - 2009. - Pp. 85-92.