На правах рукописи

Епрев Антон Сергеевич

ИССЛЕДОВАНИЕ ВЛИЯНИЯ РАЗРЕШЕНИЯ ЛЕКСИЧЕСКОЙ МНОГОЗНАЧНОСТИ С ПОМОЩЬЮ КОНТЕКСТНЫХ ВЕКТОРОВ НА ЭФФЕКТИВНОСТЬ КАТЕГОРИЗАЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ

05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата физико-математических наук

Новосибирск – 2011

Работа выполнена в Омском государственном университете
им. Ф.М. Достоевского.

Научный руководитель:	доктор физико-математических наук Белим Сергей Викторович
Официальные оппоненты:	доктор физико-математических наук Зыкина Анна Владимировна
	кандидат физико-математических наук Батура Татьяна Викторовна
Ведущая организация:	Институт математики им. С.Л. Соболева Сибирского отделения РАН

Защита состоится 22 июня 2011 г. в 16 ч. 00 мин. на заседании диссертационного совета ДМ 003.032.01 при Институте систем информатики им. А.П. Ершова Сибирского отделения РАН, расположенном по адресу: 630090, г. Новосибирск, пр. Лаврентьева, д. 6.

С диссертацией можно ознакомиться в библиотеке Института систем информатики Сибирского отделения РАН.

Автореферат разослан 20 мая 2011 г.

Ученый секретарь
диссертационного совета,

к. ф.-м. н. Мурзин Ф.А.

Общая характеристика работы

Актуальность работы. Объем накапливаемой и обрабатываемой информации постоянно увеличивается, что приводит к сложности ориентирования в информационных ресурсах, и делает задачу категоризации текстовых документов все более актуальной. Использование классификаторов позволяет ограничить поиск необходимой информации относительно небольшим подмножеством документов. Так, например, в «автоматизированной системе тематического анализа информации» (Васенин В. А. и др., 2009) классификатор используется для фильтрации результатов поиска, что повышает релевантность поисковой выдачи. Помимо сужения области поиска в поисковых системах, задача категоризации имеет практическое применение в следующих областях: фильтрация спама, составление тематических каталогов, контекстная реклама, системы электронного документооборота, снятие омонимии в автоматическом переводе текстов.

Категоризация текстовых документов является задачей автоматического отнесения документа к одной или нескольким категориям на основании содержания документа. Существуют различные модели и методы категоризации текстов — деревья решений, метод наименьших квадратов, адаптивные линейные классификаторы, метод ближайших соседей, метод опорных векторов и другие (Sebastiani F., 2002).

В последнее время активно разрабатываются способы интеграции различных баз знаний и ресурсов в методы категоризации текстовых документов с целью получения высоких результатов категоризации. Большой интерес представляет использование семантических ресурсов, таких как WordNet или Wikipedia.

WordNet — это семантический словарь английского языка, базовой словарной единицей которого является синонимический ряд, так называемый

«синсет», объединяющий слова со схожим значением. Синсеты связаны между собой различными семантическими отношениями. Также существуют реализации для других языков, ведутся разработки WordNet для русского языка.

Большинство методов категоризации основывается на использовании простой векторной модели описания документов, в которой признаками документов являются базовые формы слов. Использование слов в качестве признаков имеет ряд недостатков: словосочетания, такие как «European Union», разделяются на отдельные слова и обрабатываются независимо; слова, являющиеся синонимами, используются как самостоятельные признаки; многозначные слова рассматриваются как обычные признаки, в то время как они могут иметь несколько различных значений. В работе (Gonzalo J. et al., 1998) отмечается, что использование в качестве признаков документов значений слов, представленных синсетами, может приводить к улучшению качества категоризации на 28%. Такие результаты были получены на коллекции документов, где устранение лексической многозначности слов было выполнено *вручную*. Согласно результатам исследования, эффективность категоризации при использовании методов автоматического разрешения лексической многозначности, доля ошибок которых составляет менее 10%, сопоставима с эффективностью категоризации для размеченного вручную текста. Увеличение доли ошибок разрешения лексической многозначности с 10% до 30% приводит к резкому спаду эффективности категоризации, а для методов с ошибкой 30–60% использование в качестве признаков синсетов не приводит к заметному приросту эффективности категоризации.

Существует несколько публикаций, в которых сравниваются эффективности категоризации с использованием слов и синсетов WordNet, полученных с помощью различных методов *автоматического* разрешения лексической многозначности. В системе автоматической категоризации документов на базе метода k–ближайших соседей (Ferretti E. et al., 2003) использование син-

сетов в качестве признаков, полученных с помощью метода, базирующегося на использовании скрытой модели Маркова, приводит к росту эффективности категоризации на 2%. В работе (Bloehdorn S. et al., 2004) проводилось сравнение алгоритма категоризации «AdaBoost» на нескольких коллекциях документов, а для устранения лексической многозначности слов применялся метод, суть которого заключается в выборе того синсета, слова которого в документе встречаются чаще остальных. Использование данного метода позволяет повысить эффективность категоризации на 1%.

В работе (Patwardhan S. et al., 2006) описывается метод оценки семантической близости синсетов с помощью контекстных векторов, использующий информацию о совместной встречаемости слов в тексте. Оценка эффективности этого метода проводилась на нескольких наборах слов. Данный метод показывает лучшие результаты среди других методов оценки семантической близости слов на базе ресурса WordNet. Однако, практическое применение данного метода для устранения лексической многозначности не исследовалось.

Актуальность исследования обуславливается практической значимостью систем автоматической категоризации текстовых документов, в которых в качестве признаков используются значения слов, представленные синсетами WordNet.

Цели диссертационной работы:

1. Разработать и реализовать алгоритм разрешения лексической многозначности слов с помощью контекстных векторов на базе ресурса WordNet.
2. Реализовать программный комплекс автоматической категоризации текстовых документов с использованием синсетов WordNet в качестве признаков документов.
3. Исследовать применимость разработанного алгоритма разрешения лек-

сической многозначности к различным коллекциям документов с помощью оценки его влияния на эффективность категоризации.

Научная новизна исследования состоит в следующем:

1. Разработан алгоритм разрешения лексической многозначности слов, в котором используются контекстные векторы для оценки семантической близости синсетов с контекстом.
2. Реализован программный комплекс автоматической категоризации текстовых документов, в котором используются синсеты WordNet в качестве признаков документов и контекстные векторы для разрешения лексической многозначности.

Практическая значимость заключается в формировании нового инструмента, позволяющего повысить эффективность категоризации текстовых документов.

Полученные в диссертации результаты могут быть использованы в существующих информационных системах для повышения релевантности результатов поиска, в системах электронного документооборота для тематической категоризации документов, и представляют научный интерес для специалистов в области информационного поиска и машинного обучения.

Основные положения, выносимые на защиту:

1. Алгоритм разрешения лексической многозначности слов, в котором используются контекстные векторы для оценки семантической близости синсетов с контекстом.
2. Алгоритм обработки текстовых документов, позволяющий выделять в тексте словосочетания произвольной длины, для которых существуют синсеты WordNet.

3. Повышение качества категоризации неспециализированных текстов при использовании в качестве признаков документов синсетов WordNet, полученных с помощью разработанного алгоритма разрешения лексической многозначности.
4. Влияние на качество категоризации тематики корпуса для построения пространства слов, в котором представляются контекстные векторы.

Апробация работы. Основные результаты диссертации докладывались на следующих конференциях и семинарах: XVIII всероссийский семинар «Нейроинформатика, ее приложения и анализ данных», г. Красноярск, Академгородок, 2010; II международная научно–практическая конференция «Прогрессивные технологии и перспективы развития», г. Тамбов, 2010; II международная заочная научно–практическая конференция «Современные направления научных исследований», 2010; межвузовская научно–практическая конференция «Информационные технологии и автоматизация управления», г. Омск, 2009; научный семинар кафедры информационной безопасности факультета компьютерных наук ОмГУ им. Ф. М. Достоевского, г. Омск, 2010.

Публикации. Материалы диссертации опубликованы в 10 печатных работах, из них 2 статьи в журналах из списка, рекомендованного ВАК.

Личный вклад автора. Содержание диссертации и основные положения, выносимые на защиту, отражают персональный вклад автора в опубликованные работы. Все представленные в диссертации результаты получены лично автором.

Структура и объем диссертации. Диссертация состоит из введения, трех основных глав, заключения и библиографии. Общий объем диссертации 118 страниц, содержит 16 рисунков и 18 таблиц. Библиография включает 112 наименований.

Содержание работы

Во **введении** обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, показана практическая значимость полученных результатов и представлены выносимые на защиту научные положения.

Первая глава посвящена обзору методов автоматической категоризации текстовых документов.

Дается определение категоризации текстовых документов как задачи автоматического отнесения документа к одной или нескольким категориям на основании содержания документа. Приводится формальная постановка задачи.

Задача категоризации текстовых документов рассматривается как задача аппроксимации неизвестной функции $\Phi : D \times C \to \{0, 1\}$, определяющей каким образом документы должны быть классифицированы, через функцию $\hat{\Phi} : D \times C \to \{0, 1\}$, именуемую классификатором, где $C = \{c_1, \ldots, c_{|C|}\}$ — множество возможных категорий, а $D = \{d_1, \ldots d_{|D|}\}$ — множество документов.

Выделяется особый вид классификаторов — *бинарные*, множество категорий которых состоит из двух элементов (c_i и его дополнения $\bar{c}_i$). Бинарный классификатор для $\{c_i, \bar{c}_i\}$ определяется функцией $\hat{\Phi}_i : D \to \{0, 1\}$, которая является аппроксимацией неизвестной функции $\Phi_i : D \to \{0, 1\}$.

Нахождение классификатора для множества категорий $C = \{c_1, \ldots, c_{|C|}\}$ рассматривается как поиск $|C|$ бинарных классификаторов $\{c_i, \bar{c}_i\}$, где $i = 1, \ldots, |C|$.

Формулируется задача автоматической категоризации текстовых документов, которая включает в себя следующие этапы:

1. *Индексирование документов.* Документы на естественном языке необхо-

димо преобразовать в удобную для машинной обработки форму. В процессе индексирования происходит выделение признаков из документов. На этом этапе определяется числовая модель представления документа.

2. *Построение классификатора.* На этом этапе применяются различные методы машинного обучения. Классификатор для категории c_i автоматически создается в процессе обучения, при котором просматривается множество документов с заранее определенными категориями c_i или $\bar{c}_i$ и подбираются такие характеристики классификатора, чтобы новый документ, отнесенный к категории c_i, соответствовал им.
3. *Вычисление эффективности классификатора.* Эффективность классификатора является качественной оценкой результатов его работы на некотором множестве документов, для которого известны значения Φ. Эффективность используется для сравнения различных методов категоризации.

Для каждого этапа приводится описание применяемых методов и используемых моделей представления документов.

Рассматриваются ансамбли из классификаторов, которые позволяют повысить точность категоризации с помощью построения k классификаторов $\hat{\Phi}_1, \ldots, \hat{\Phi}_k$ и объединении результатов их работы.

Вторая глава посвящена методам разрешения лексической многозначности на базе WordNet и разработанному алгоритму категоризации текстовых документов, в котором используются синсеты в качестве признаков документов и контекстные векторы для устранения лексической многозначности слов.

Разрешение лексической многозначности (Word Sense Disambiguation) — это задача автоматического выбора значения многозначного слова или фразы из множества их значений в зависимости от контекста, в котором данное слово или словосочетание находятся.

Рассматриваются различные методы разрешения лексической многозначности на базе WordNet. Особое внимание уделяется методу оценки семантической близости синсетов с помощью контекстных векторов.

В определении значений слов существенную роль играет контекст. Одно и то же значение слова, как правило, употребляется в одинаковом контексте. Контекстные векторы широко используются в информационном поиске и в задачах обработки естественного языка. Контекстный вектор (первого порядка) $\vec{w}$ указывает на все слова, вместе с которыми слово w встречается в тексте. Векторы, сформированные из суммы контекстных векторов (контекстные векторы второго порядка), используются для представления значений слов.

Чтобы построить контекстные векторы первого порядка, необходимо определить пространство слов W, обработав некоторый корпус текстов. В качестве такого корпуса используется объединение дефиниций синсетов WordNet. Полученный корпус содержит приблизительно 860 тысяч слов, из которых около 40 тысяч являются уникальными. Также исключаются из рассмотрения редко встречающиеся и стоп–слова, что позволяет сократить размерность пространства слов W до порядка 20 тысяч.

После построения контекстных векторов первого порядка, производится вычисление векторов дефиниций синсетов WordNet. Вектор дефиниции определяется как результат сложения контекстных векторов первого порядка слов, входящих в определение синсета. Например, дефиницией одного из значений слова «fork» является выражение «cutlery used to serve and eat food». Вектор дефиниции для него представляет собой результат сложения контекстных векторов первого порядка слов «cutlery», «serve», «eat» и «food».

Разработанный алгоритм разрешения лексической многозначности основывается на предположении, что два контекстных вектора второго порядка, расположенные близко друг к другу, скорее всего обозначают одно и тоже

значение слова. Таким образом, определение значения многозначного слова w в некотором предложении заключается в следующем:

1. Вычисляется вектор для контекста слова w, который является суммой контекстных векторов первого порядка слов, находящихся на расстоянии в несколько позиций слева и справа от w в предложении.
2. Производится оценка семантической близости всех возможных значений слова w с контекстом. Для каждого синсета слова w вычисляется косинус угла между вектором его дефиниции и вектором контекста.
3. Самый близкий к контексту синсет выбирается в качестве значения слова w.

Приводится пошаговое описание алгоритма категоризации документов на базе WordNet с использованием контекстных векторов для разрешения лексической многозначности слов. Алгоритм категоризации основан на методе k–ближайших соседей (k–NN) и использует в качестве признаков документов синсеты WordNet.

Построение классификатора начинается с индексирования документов обучающей коллекции $\mathcal{L}$. На этом этапе происходит морфологический разбор слов, встречающихся в документах, поиск словосочетаний и устранение лексической многозначности. На выходе каждый документ описывается множеством признаков, представленных синсетами WordNet.

Затем осуществляется процедура уменьшения размерности пространства признаков с использованием функции полезности на базе критерия χ^2 (Yang Y. et al., 1997). Функция полезности $f(t_k, c_i)$ характеризует значимость признака t_k в некотором документе для категории c_i:

$$f(t_k, c_i) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)},$$

где N — количество документов в обучающей коллекции, A — количество

документов, в которых встречается t_k и которые определены в c_i, B — количество документов, в которых встречается t_k и которые не определны в c_i, C — количество документов, в которых не встречается t_k и которые определены в c_i, D — количество документов, в которых не встречается t_k и которые не определены в c_i.

Чтобы вычислить значимость признака t_k для всех категорий C, необходимо найти максимальное значение $f(t_k) = max_{i=1}^{|C|} f(t_k, c_i)$.

После уменьшения размерности пространства признаков осуществляется вычисление весовых коэффициентов признаков для документов обучающей коллекции. Для взвешивания используется один из вариантов статистических весовых функции «$tf \cdot idf$» (Salton G. et al., 1998):

$$\omega_{ij} = \frac{tf_{ij} \cdot idf_i}{\sqrt{\sum_k (tf_{kj} \cdot idf_k)^2}},$$

где ω_{ij} — вес i-го признака в документе d_j, tf_{ij} — частота встречаемости i-го признака в рассматриваемом документе, idf_i — логарифм отношения количества документов в коллекции к количеству документов, в которых встречается i-ый признак. Веса, вычисленные по этой формуле, нормализованы таким образом, что сумма квадратов весов каждого документа равна единице. Документ d_j после взвешивания представляется вектором $\vec{d_j} = \langle \omega_{1j}, \ldots, \omega_{|T|j} \rangle$.

На этом построение классификатора заканчивается. Категоризация новых документов осуществляется следующим образом. Документ d_j, поступающий в систему, проходит через тот же механизм индексирования с последующим взвешиванием признаков. Для того чтобы определить категории, соответствующие документу d_j, классификатор выполняет следующие действия:

1. Документ d_j сравнивается со всеми документами из обучающей коллекции $\mathcal{L}$. Для каждого $d_z \in \mathcal{L}$ вычисляется расстояние $\rho(d_j, d_z)$ — значение

косинуса угла между векторами $\vec{d_j}$ и $\vec{d_z}$.

2. Далее из обучающей коллекции выбираются k ближайших к d_j документов.
3. Определение категорий документа d_j осуществляется выбором наиболее встречающихся категорий среди k ближайших к d_j документов, т. е. документ относится к категориям, частота встречаемости cf_i которых больше некоторого τ.

Значение τ было определено в ходе экспериментов и равняется $0.8 \cdot \max_i cf_i$. В качестве значения k используется 30.

Третья глава посвящена используемым для вычисления эффективности разработанного классификатора корпусам текстов и результатам экспериментов.

Рассматривается программная реализация классификатора. Программный комплекс автоматической категоризации реализован на языке программирования Erlang с использованием открытой платформы OTP.

Приводится описание корпусов текстов «Reuters–21578» и «Reuters Corpus Volume 1», которые использовались для сравнения эффективности построенных классификаторов. Для коллекции «Reuters–21578» существуют фиксированные разбиения на обучающее и тестирующее множества. Построение классификатора и оценка его эффективности проводилась с использованием разбиения «ModApt». Это разбиение задает 90 категорий, 9603 документа содержатся в обучающем наборе и 3299 документов в тестирующем.

Для коллеции «Reuters Corpus Volume 1» не предусмотрены стандартные разбиения. Для экспериментов были выбраны 10 разносторонних категорий: международные отношения; катастрофы и бедствия; искусство, культура и сфера развлечений; мода; здоровье; религия; наука и технологии; спорт; путешествия и туризм и погода. Из всей коллекции были отобраны 5923 доку-

мента, определенных в одну или несколько вышеперечисленных категорий, и разделены на два множества. Обучающий набор содержит 3532 документа, тестовый набор — 1761.

Для исследования влияния разработанного алгоритма разрешения лексической многозначности на эффективность категоризации текстовых документов были проведены эксперименты с разработанным k–NN классификатором и классификатором SVM^{light} (Joachims J., 1999). Для каждой коллекции документов сначала проводилось вычисление эффективности категоризации, когда в качестве признаков документов выступали базовые формы слов (т. е. без использования WSD), а затем — синсеты (и использованием WSD).

Оценка эффективности категоризации рассматривается как комбинация точности p и полноты r. Точность — это доля верно классифицированных документов, а полнота — отношение верно классифицированных документов к общему количеству документов, которые должны были быть классифицированы. Точность и полнота вычисляются для каждой категории индивидуально, затем происходит их микро– и макроусреднение. Чем больше точность и полнота, тем качественнее результаты категоризации.

В таблицах 1 и 2 приведены результаты экспериментов для k–NN классификатора на коллекциях документов «Reuters–21578» и «Reuters Corpus Volume 1», а в таблицах 3 и 4 — для SVM^{light}.

Таблица 1. Эффективность k–NN классификатора на коллекции «Reuters–21578»

Классификатор	Микро p	Микро r	Макро p	Макро r
Без использования WSD	.8340	**.7727**	.8939	**.2993**
С использованием WSD	**.8380**	.7664	**.9187**	.2869

Результаты экспериментов показывают, что использование синсетов в качестве признаков документов, полученных с помощью разработанного алго-

Таблица 2. Эффективность k–NN классификатора на коллекции «Reuters Corpus Volume 1»

Классификатор	Микро p	Микро r	Макро p	Макро r
Без использования WSD	.8499	.8569	.8611	.8231
С использованием WSD	**.8672**	**.8605**	**.8850**	**.8287**

Таблица 3. Эффективность SVM классификатора на коллекции «Reuters–21578»

Классификатор	Микро p	Микро r	Макро p	Макро r
Без использования WSD	**.9481**	**.7911**	**.9591**	.3852
С использованием WSD	.9294	.7807	.9066	**.9079**

ритма разрешения лексической многозначности, позволяет повысить эффективность категоризации для коллекции неспециализированных текстов. В то же время на корпусе текстов узкой направленности (коллекция «Reuters–21578» содержит статьи финансового характера) для k–NN классификатора происходит увеличение точности в ущерб полноте, а для SVM классификатора происходит увеличение макроусредненной полноты при небольшом снижении остальных значений.

На корпусе «Reuters Corpus Volume 1» было проведено исследование зависимости эффективности категоризации от длины документов. Для этого документы тестирующей коллекции были разбиты на 5 групп по количеству символов. Затем для каждой группы были вычислены значения эффективности категоризации без/с использованием WSD. На рисунке 1 показана зависимость микроусредненного значения меры $F_1 = 2pr/(p + r)$ для k–NN классификатора от длины документов. С ростом длины документов происходит увеличение прироста эффективности при использовании WSD до 4%.

В **заключении** сформулированы выводы и основные результаты работы.

Таблица 4. Эффективность SVM классификатора на коллекции «Reuters Corpus Volume 1»

Классификатор	Микро p	Микро r	Макро p	Макро r
Без использования WSD	**.9580**	.8646	**.9600**	.8402
С использованием WSD	.9518	**.8708**	.9533	**.8536**

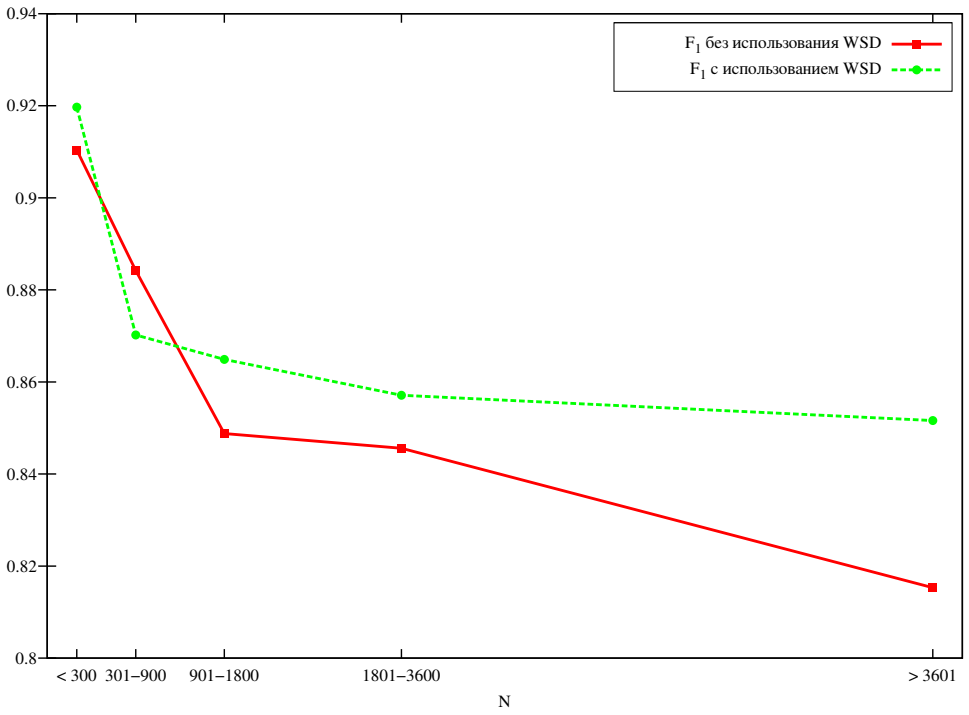

Рис. 1. Зависимость микроусредненного значения меры эффективности F_1 для k–NN классификатора без/с использования WSD от длины документов N

Основные результаты:

1. Разработан и реализован алгоритм разрешения лексической многозначности слов, в котором используются контекстные векторы для оценки семантической близости синсетов с контекстом.
2. Разработан и реализован алгоритм обработки текстовых документов, позволяющий выделять в тексте словосочетания произвольной длины, для которых существуют синсеты WordNet.
3. Разработан и реализован алгоритм категоризации текстовых документов на базе метода k–ближайших соседей, в котором синсеты использу-

ются в качестве признаков документов.

4. Показано, что использование в текстовом классификаторе в качестве признаков документов синсетов WordNet, полученных с помощью разработанного алгоритма, позволяет повысить эффективность категоризации неспециализированных текстов.
5. Выявлено, что тематика корпуса текстов для построения пространства слов, в котором представляются контекстные векторы, оказывает влияние на качество категоризации.
6. Показано, что эффективность категоризации зависит от размера обрабатываемых документов. Увеличение длины документов сопровождается снижением качества категоризации. Но в тоже время использование разработанного алгоритма разрешения лексической многозначности позволяет добиться улучшения качества категоризации длинных документов.

Основные публикации по теме диссертации

Журналы из списка, рекомендованного ВАК:

1. А.С. Епрев. Применение разрешения лексической многозначности в классификации текстовых документов // Наука и образование. 2010. № 10. С. 1–4.
2. А.С. Епрев. Применение контекстных векторов в классификации текстовых документов // Журнал радиоэлектроники. 2010. № 10. С. 1–7.

Остальные публикации:

3. А.С. Епрев. Обзор методов классификации текстов // Проблемы обработки и защиты информации. Книга 2. Анализ графической и текстовой

информации. Коллективная монография / Под общей ред. д. ф.-м. н. С.В. Белима. – Омск: ООО «Полиграфический центр КАН». 2010. С. 5–28.

4. А.С. Епрев. Применение баз знаний в задачах классификации текстов // Проблемы обработки и защиты информации. Книга 2. Анализ графической и текстовой информации. Коллективная монография / Под общей ред. д. ф.-м. н. С.В. Белима. – Омск: ООО «Полиграфический центр КАН». 2010. С. 29–42.
5. А. С. Епрев. Тематическая классификация документов по степени близости термов // Математические структуры и моделирование. 2009. № 20. С. 93–96.
6. А. С. Епрев. Автоматическая классификация текстовых документов // Математические структуры и моделирование. 2010. № 21. С. 65–81.
7. А. С. Епрев. Использование WordNet в k-NN классификаторе // Материалы XVIII Всероссийского семинара «Нейроинформатика, ее приложения и анализ данных». Красноярск, 2010. С. 68–72.
8. А. С. Епрев. Методы разрешения лексической многозначности на базе WordNet // Материалы II международной заочной научно–практической конференции «Современные направления научных исследований». Екатеринбург, 2010. С. 85–86.
9. А. С. Епрев. Интеграция семантического словаря WordNet в текстовый классификатор // Материалы II международной научно–практической конференции «Прогрессивные технологии и перспективы развития». Тамбов, 2010. С. 25–26.
10. А. С. Епрев. Тематическая классификация документов // Материалы межвузовской научно–практической конференции «Информационные технологии и автоматизация управления». Омск, 2009. С. 129.

Епрев Антон Сергеевич

ИССЛЕДОВАНИЕ ВЛИЯНИЯ РАЗРЕШЕНИЯ ЛЕКСИЧЕСКОЙ МНОГОЗНАЧНОСТИ С ПОМОЩЬЮ КОНТЕКСТНЫХ ВЕКТОРОВ НА ЭФФЕКТИВНОСТЬ КАТЕГОРИЗАЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ

Автореф. дисс. на соискание ученой степени кандидата физико-математических наук.
Подписано в печать 16.05.2011. Заказ № 1050. Формат 84x54 1/16. Усл. печ. л. 1. Тираж 100 экз.
Отпечатано в ООО «БЛАНКОМ», г. Омск, пр. К. Маркса, 18 корп. 8.