Метод автоматической классификации коротких текстовых сообщений[1]

Э. Мбайкоджи, А.А. Драль, И.В. Соченков

Аннотация. В статье представлены результаты исследования в области классификации коротких текстовых документов. Проанализированы методы классификации на основе анализа распределения лексических дескрипторов естественного языка. Описан метод оценки информационной значимости в текстах естественного языка. Представлен метод классификации текстовых документов на основе характеристики тематической значимости.

Ключевые слова: классификация коротких текстовых документов, классификация по метаданным, мультиномиальная модель, метод опорных векторов, TF, IDF, характеристика тематической значимости.

Введение

Информационные системы являются неотъемлемой частью повседневной человеческой деятельности. Одним из крупнейших информационных ресурсов универсального назначения является Интернет, основанный на технологиях гипертекста. Объёмы доступных информационных источников в Интернете оцениваются по данным на 2012 в 628 миллионов сайтов [24]. Большое количество ресурсов снижает доступность информации: в информационном изобилии становится сложным отделить нужное в данный момент от всего остального. Эта проблема существует как в Интернете в целом, так и в отдельных его сегментах: лентах новостей, социальных сетях, блогах и приобретающих всё большую популярность микроблогах.

Традиционным решением проблемы доступа к информации в Интернете являются глобальные поисковые системы (Яндекс, Google). Однако их функциональность ограничивается классическим поиском. При этом в Интернете ежедневно создаётся большое количество новых web-страниц: по данным на 2011 год ежедневно создается 150 тысяч новых страниц [22], содержащих новости, сообщения, комментарии и мнения пользователей. Эта информация имеет неструктурированный характер: как сайты, так и их содержание различаются по структуре, типу, тематике и другим критериям. В условиях развивающегося Интернета существует потребность в более глубокой обработке и анализе накопленной и вновь создаваемой информации, например, с целью мониторинга общественного мнения, оценки популярности товаров и услуг, создания подборок документов некоторой предметной области и др.

Эти задачи решаются с помощью информационно-аналитических систем с применением методов анализа естественно-языковых текстов, классификации и кластеризации информации. С помощью этих методов возможно определить, к какой тематике принадлежит тот или иной текст, каковы содержательные особенности текста, как тексты связаны между собой. Поэтому информационно-аналитические системы востребованы при обработке неструк-

[1] Работа выполнена при финансовой поддержке Минобрнауки России (Госконтракт № 14.514.11.4024) в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы».

турированной информации, представленной в Интернете, а также в научно-образовательных сетях, электронных библиотеках, системах электронного документооборота.

Важной частью информационно-аналитических систем является тематический рубрикатор. Рубрики представляют собой тематически однородные множества документов. При этом множество рубрик задаётся априорно экспертами. На этапе обработки документов в информационно-аналитической системе функционирует автоматический классификатор – программа, определяющая тематику документов и осуществляющая их отнесение к рубрикам.

Настоящая работа посвящена экспериментальной проверке ряда общеизвестных методов классификации информации применительно к задаче тематической классификации коротких текстовых сообщений. К коротким сообщениям относят текстовые документы, средняя длина которых составляет менее 2000-3000 символов [8], например, сообщения в социальной сети Твиттер (Twitter), средняя длина которых составляет 78 символов [15].

В настоящей работе авторами предложен метод автоматической классификации текстов на основе характеристики тематической значимости слов естественного языка. Целью исследований являлось создание нового метода автоматической классификации текстовой информации, применимого в информационно-аналитических системах, обладающего высокой вычислительной эффективностью и высоким качеством классификации при низкой ресурсоёмкости и невысокой вычислительной сложности.

1. Математическая модель классификатора текстов

В этом разделе формализуется модель автоматического классификатора текстов в виде алгебраической системы, в рамках которой впоследствии будут сформулированы и описаны методы классификации, исследуемые в настоящей работе.

Общую модель классификатора текстов представим в виде следующей системы:

$$R = < T, C, F, R_C^F, f >, \qquad (1)$$

где: T – коллекция текстов, C – непустое множество тематических рубрик, F – непустое множество описаний тематических рубрик (каждое описание содержит данные, необходимые для классификации, например, списки ключевых слов и их значимость), R_C^F – отношение на $C \times F$, соотносящее тематические рубрики и соответствующие им описания, f – операция классификации – отображение $T \to 2^C$, такое что $f(t) = \sigma$, где t - текст из T, а $\sigma \in 2^C$ – элемент множества всех подмножеств C, т.е. множество тематических рубрик из C. Таким образом, отображение f позволяет каждому документу множества T поставить в соответствие некоторую тематическую рубрику из C.

Будем также полагать, что:

1. Отношение R_C^F функционально, т.е. обладает свойством: $\forall c \in C \ \exists!$, $\varphi \in F : (c, \varphi) \in R_C^F$ то есть каждой тематической рубрике соответствует единственное описание. Обратное необязательно.

2. Каждое описание тематической рубрики содержит набор признаков, используемых операцией классификации, а также их значения.

3. Для каждой тематической рубрики c определим $D(c)$ – суть множество соотнесённых с ней документов (априорно или с помощью операции классификации), т.е. $D(c) = \{t \in T | c \in f(t)\}$. Рубрика $c_i \subset c_j$ (c_i вложена в c_j) по определению, если $D(c_i) \subset D(c_j)$.

4. Можно показать, что C является верхней полурешеткой, т.е. существует единственный элемент $\sigma_{max} \in 2^C$, такой что для всякого $\sigma \in 2^C$ $\sigma \in \sigma_{max}$. Т.е. σ_{max} корневая рубрика, содержащая в себе все остальные классы.

5. Результат определения тематической принадлежности документа – классификация документа – $t \in T$ – есть σ - множество тематических рубрик, которым соответствует документ. Невозможность определить тематическую принадлежность документа $t \in T$ означает, что либо это множество пусто либо $t \in \sigma_{max}$.

В задачах иерархической классификации на множестве тематических рубрик C задаётся отношение, определяющее априорную вложенность рубрик друг в друга. В настоящей работе

мы ограничимся рассмотрением задачи плоской классификации, при которой все рубрики вложены в корневую, а иная вложенность отсутствует.

Система R представляет собой модель автоматического классификатора документов текстовых коллекций на естественном языке. Построение классификатора подразумевает частичное или полное формирование C, F, R_C^F, f на основе некоторых априорных данных [1]. На практике это означает, что экспертом формируется иерархия тематических рубрик. Описания тематических рубрик формируются либо вручную (например, в виде правил отнесения документов к тематическим рубрикам на основе некоторых признаков) либо автоматически с применением методов машинного обучения. В качестве обучающего множества выступает набор документов, заранее соотнесённых с категориями T_0 на основе экспертных оценок.

Для дальнейшего рассмотрения текстов как объектов классификации и выделения значимых признаков классификации введём понятие лексического дескриптора. Под лексическим дескриптором (ЛД) естественного языка (ЕЯ) будем понимать [3,9,12]:

– отдельные лексемы как совокупности парадигматических форм (словоформ) одного слова, т.е. разные формы одного и того же слова не различаются;

– словосочетания в канонических формах (главное слово приведено к словарной форме, а форма зависимых слов подчинена управлению главного слова) безотносительно различных форм.

Текст документа t характеризуется множеством лексических дескрипторов, содержащихся в этом тексте: $\phi_t = \{w_i\}_{i=1}^{S_t}$, S_t – общее количество лексических дескрипторов в тексте. L_t – множество вхождений отдельных слов в текст.

Для ЛД текста документа, как правило, определены также дополнительные числовые характеристики (например, частота встречаемости, значимость, вес в тексте и т.п.). В силу этого документ характеризуется вектором значений признаков, где в качестве пространства признаков выступает φ_t, а в качестве координат – конкретные числовые значения.

Таким образом, решение задачи классификации состоит в выборе для документа t подходящей тематической рубрики (или нескольких рубрик) либо отнесение к корневой рубрике (отказ от классификации) в случае открытой классификации.

2. Постановка задачи исследования

В настоящей работе решается задача сравнительной экспериментальной оценки качества работы нескольких методов автоматической классификации коротких текстовых сообщений. В качестве исходных данных для классификации выступают текстовые документы на русском языке – рекламные объявления. Решаемая задача классификации является закрытой, т.е. каждое рекламное объявление относится как минимум к одной тематической рубрике (не считая корневую). Рубрикатор является плоским. Для формирования описаний рубрик имеется обучающая выборка рекламных объявлений, экспертно соотнесённых с рубриками. При этом описания рубрик формируются автоматически с применением методов машинного обучения.

Для оценки степени принадлежности документа тематической рубрике используются следующие метрики:

1) расстояние Евклида [17];
2) мера Хеллингера [21];
3) косинусная мера сходства [18];
4) метрика близости в рамках мультинамиальной модели [7];
5) характеристика тематической значимости текста (ХТЗ) [5, 14], и её модификация, предложенная авторами.

Для расчёта значений признаков классификации и определения расстояний «документ-рубрика» с помощью вышеуказанных метрик используются следующие формулы и определения.

Относительная частота встречаемости ЛД в тексте t:

$$TF(w,t) = \frac{k(w, L_t)}{S(L_t)}, \qquad (2)$$

где $k(w, L_t)$ – количество вхождений ЛД w в тексте t, $S(L_t)$ – общее количество вхождений ЛД в текст t.

Относительная частота встречаемости ЛД в текстах рубрики c:

$$TF(w,c) = \frac{k(w, L_c)}{S(L_c)}, \quad (3)$$

где $k(w, L_c)$ – количество вхождений ЛД w в составной текст $L_c = \bigcup_{t \in D(c)} L_t$ тематической рубрики c (полученный объединением текстов всех относящихся к ней документов), $S(L_c)$ – общее количество вхождений ЛД в составной текст рубрики c.

Инверсная частота ЛД w в произвольном множестве текстов τ:

$$IDF(w,\tau) = \log_2 \frac{|\tau|}{m(w,\tau)} \quad (4)$$

где $m(w,\tau) = \left|\{t \in \tau \mid w \in \varphi_t\}\right|$ – число документов в τ, содержащих лексический элемент w; $|\tau|$ – общее число документов в τ. В нормированном на 1 варианте величина $IDF(w, \tau)$ имеет вид:

$$IDF_N(w,\tau) = \frac{\log_2 \frac{|\tau|}{m(w,\tau)}}{\log_2 |\tau|} = \log_{|\tau|} \frac{|\tau|}{m(w,\tau)} \quad (5)$$

Эта величина определяет количество информации, связанное с наступлением события «при случайном выборе документа из τ встретить в этом документе хотя бы одно вхождение ЛД w».

Для вычислительной корректности расчёта величины (5) положим, что если $m(w, \tau) = 0$, то $IDF_N(w, \tau) = 1$ по определению.

Вес или значимость ЛД w текста t в общетематической коллекции текстов T определяется нормированной на 1 неотрицательной величиной:

$$TFIDF_N(w,t) = TF(w,t) \cdot IDF_N(w,T) \quad (6)$$

Вес или значимость ЛД w в текстах рубрики c определяется нормированной на 1 неотрицательной величиной:

$$TFIDF_N(w,c) = TF(w,c) \cdot IDF_N(w, D(c)) \quad (7)$$

С учетом введенных определений расстояние для определения близости «документ-рубрика» в метрике Евклида принимает вид:

$$Euclidean(t,c) = \sqrt{\sum_{w \in \varphi_t \cap \varphi_c} (TFIDF_N(w,c) - TFIDF_N(w,t))^2} \quad (8)$$

Косинусная мера для определения близости «документ-рубрика»:

$$Cos(t,c) = \frac{\sum_{w \in \varphi_c \cap \varphi_t} TFIDF_N(w,c) \cdot TFIDF_N(w,t)}{\sqrt{\sum_{w \in \varphi_c} TFIDF_N(w,c)} \cdot \sqrt{\sum_{w \in \varphi_t} TFIDF_N(w,t)}} \quad (9)$$

Мера Хеллингера для определения близости «документ-рубрика»:

$$Hellinger(t,c) = 1 - \sum_{w \in \phi_c \cap \phi_t} \sqrt{TFIDF_N(w,c) \cdot TFIDF_N(w,t)} \quad (10)$$

Мера определения близости «документ-рубрика» в рамках мультиномиальной модели представления текста:

$$Multinomial(t,c) = \prod_{w \in \varphi_c \cap \varphi_t} (TFIDF_N(w,c))^{TFIDF_N(w,t)} \quad (11)$$

В качестве решающего правила классификации для метрики Евклида и меры Хеллингера используется принцип минимума: документ относится к той рубрике, для которой метрика принимает минимальное значение. Напротив, для косинусной меры и меры близости в рамках мультиномиальной модели документ относился к рубрике, для которой метрика достигла максимального значения.

Для решения задачи классификации текстовых документов наряду с вышеперечисленными методами в настоящем исследовании используется метод опорных векторов – SVM [23]. Это один из стандартных методов классификации объектов, представимых в виде численных векторов. При этом процесс классификации сводится к нахождению гиперплоскости, которая разделяет классифицируемые объекты (разделяющая гиперплоскость). При применении метода опорных векторов для каждого

классифицируемого объекта рассчитываются значения близости «документ- рубрика» для каждой тематической рубрики по выбранной метрике. Таким образом, для каждого документа получается новое признаковое описание: числовой вектор размерности $|C|$ - мощность непустого множества тематических рубрик, коэффициентами которого являются расстояния «документ-рубрика». Полученные вектора признаков подаются на вход алгоритму классификации на базе опорных векторов. В экспериментах с методом опорных векторов также были использованы комбинации нескольких метрик: вектора размерности $k*|C|$, полученные конкатенацией k векторов расстояний «документ-рубрика», рассчитанных с помощью различных метрик.

При решении задачи классификации авторами также исследована ХТЗ текста и предложена её модификация, описанию которой посвящён следующий раздел настоящей статьи.

3. Характеристика тематической значимости

В этом разделе рассматривается метод расчёта значений признаков для классификации, базирующийся на понятии информационной значимости ЛД текста на ЕЯ, и автоматический классификатор на основе этих признаков.

Следуя работам [11, 20], рассмотрим величины, заданные формулам (2) – (6), для оценки информационной значимости ЛД в текстах на ЕЯ, относящихся к некоторой коллекции. Величину IDF для некоторого ЛД w можно рассматривать как информацию, которую несёт в себе факт наличия этого лексический элемента в произвольно выбранном тексте коллекции T. Вес или значимость ЛД w текста t в общетематической коллекции текстов T определяется нормированной на 1 неотрицательной величиной $TFIDF_N(w,t)$.

В соответствии с работой [14] рассмотрим нормированную величину тематической инверсной частоты лексического дескриптора w относительно тематической рубрики c: $IDF_N(w,c)$. Эта величина – доля информации, приходящейся на лексический элемент w, относительно лексического элемента, встречающегося в единственном документе тематической рубрики (т.е. самого редкого лексического элемента). Смысл этой величины состоит в следующем: если тематика документа известна, т.е. очерчена некоторая тематическая область, в которой вероятно встретить характерную для неё лексику, то часть информационной неопределённости относительно содержания документа снимается. Поэтому для типичных ЛД тематической рубрики имеет место неравенство:

$$IDF_N(w,D(c)) < IDF_N(w,T) \qquad (12)$$

Положим

$$\Delta I(w,c) = IDF_N(w,T) - IDF_N(w,D(c)) \qquad (13)$$

Эта величина определяет изменение информативности выбранного ЛД w некоторого документа при отнесении этого документа к классу c. Она может рассматриваться как аналог понятия условной информации [13]. Заметим, что величина, заданная формулой (13), может быть отрицательной для ЛД, редко встречающихся в документах тематической рубрики. Введём дополнительное ограничение для $\Delta I(w,c)$ следующим образом: определим величину

$$\Delta I^+(w,c) = \mathrm{X}(\Delta I(w,c)) \cdot (\Delta I(w,c)), \qquad (14)$$

где $X(z)$ – функция Хевисайда:

$$X(z) = \begin{cases} 1, z > 0 \\ 0, z \le 0 \end{cases} \qquad (15)$$

Аналогично формуле (6) для произвольного ЛД w в текстовом документе t и тематического класса c, рассмотрим величину:

$$TFtIDF_N(w,c,t) = TF(w,t) \cdot \Delta I^+(w,c), \qquad (16)$$

которую назовём **характеристикой тематической значимости (ХТЗ)** лексического дескриптора документа [14, 5]. ХТЗ удовлетворяет условию нормировки:

$$0 \le TFtIDF_N(w,c,t) \le 1, \qquad (17)$$

В силу определений, заданных формулами (6) и (16), для любого ЛД w в произвольном тексте t и любой тематической рубрики c выполняется неравенство:

$$TFtIDF_N(w,c,t) \le TFIDF_N(w,c,t). \qquad (18)$$

В качестве величины, определяющей информативность текстового документа, будем рассматривать взвешенную сумму по некото-

рому подмножеству его ЛД. Такой подход нашёл широкое применение в задачах информационного поиска при решении задачи определения релевантности запрос-документ [16, 25, 26]. **Характеристикой значимости** текста t (ХЗ) назовём величину:

$$I(t) = \sum_{w \in \varphi_t} TFIDF_N(w,t) \qquad (19)$$

В качестве оценки информационной значимости текстового документа t в предположении, что он относится к тематической рубрике c, будем рассматривать следующую величину, которую назовём **характеристикой тематической значимости текста** – ХТЗ текста:

$$RIC(c,t) = \sum_{w \in \varphi_c \cap \varphi_t} TFtIDF_N(w,c,t). \qquad (20)$$

Формула (20) оценивает, насколько изменяется суммарная информационная ценность ЛД текста документа при отнесении этого документа к тематической рубрике c. Зная тематическую принадлежность текста, мы ожидаем обнаружить характерные тематические ЛД в этом тексте – информационная неопределённость уменьшается, а изменение количества информации, связанное со снятием неопределённости, выражается формулой (20).

Введём основанную на определениях (19), (20) величину – нормированный на 1 вариант ХТЗ:

$$RIC_N(c,t) = \frac{RIC(c,t)}{I(t)}. \qquad (21)$$

По определению для любого текста t и произвольной тематической рубрики c очевидно выполнено свойство положительной определённости и нормировки на 1:

$$0 \le RIC_N(c,t) \le 1. \qquad (22)$$

Чем больше величина ХТЗ, тем ближе тематики документа и рубрики. Верно и обратное свойство: чем она меньше, тем меньше документ и рубрика похожи тематически. Это даёт возможность использовать ХТЗ в ряде простейших решающих правил классификации: выборе максимального значения для пары «документ-рубрика» или отнесении документа к рубрике при превышении некоторого порогового значения.

Далее мы модифицируем величину из определения (13) по аналогии с классическим изменением информативности из формулы information gain [19]:

$$\Delta\tilde{I}(w,c) = IDF_N(w, T \setminus D(c)) - IDF_N(w,c). \qquad (23)$$

Т.е. величина $\Delta\tilde{I}(w,c)$ представляет собой соотношение:

$$\Delta\tilde{I}(w,c) = \log_{|T \setminus D(c)|} \frac{|T \setminus D(c)|}{m(w,T) - m(w,c)} - \log_{|D(c)|} \frac{|D(c)|}{m(w,c)}. \qquad (24)$$

Эта величина также определяет изменение информативности ЛД w (при отнесении документа, содержащего этот ЛД, к рубрике c), рассчитанное другим, по сравнению с определением (13), способом: здесь уменьшаемое – количество информации, связанное с наступлением события «встретить лексический дескриптор w вне тематического класса c».

Введём обозначение по аналогии с формулой (14):

$$\Delta\tilde{I}^+(w,c) = \Delta\tilde{I}(w,c) \cdot X(\Delta\tilde{I}(w,c)). \qquad (25)$$

Величина, заданная формулой (25), удовлетворяет требованию нормировки на 1 и неотрицательна.

В качестве модификации ХТЗ рассмотрим следующую величину:

$$R\tilde{I}C_N(c,t) = \frac{\sum_{w \in \varphi_c \cap \varphi_t} TF(w,t)\Delta\tilde{I}^+(w,c)}{\sum_{w \in \varphi_t} TF(w,t) IDF_N(w, T \setminus D(c))} \qquad (26)$$

Величина, заданная формулой (26), подобна своему аналогу (формула (21)): удовлетворяет требованию нормировки на 1, неотрицательна и может использоваться в решающих правилах классификации того же вида.

Преимуществом модифицированной ХТЗ является её нечувствительность к неравномерному разбиению обучающей выборки на классы: когда в одном рубрикаторе присутствуют большие и маленькие рубрики. Зачастую неравномерное распределение обучающей выборки не означает, что на этапе работы классификатора будет сохраняться такое же соотношение между размерами рубрик. Поэтому бывает необходимо исключить влияние

размеров рубрик на признаки классификации. Что и достигается в формуле (26).

В завершение настоящего раздела предложим способ оценки значимости ЛД в тексте отдельно взятого документа как альтернативу классической *TF*. В качестве развития подхода, предложенного в работе [5] положим:

$$lTF(w,t) = \log_{S(L_t)}\left(k(w,L_t)+1\right). \qquad (27)$$

Т.е. в качестве веса ЛД w в тексте предлагается рассматривать нормированную на 1 и неотрицательную величину, представляющую собой разность между количеством информации, приходящимся на наиболее редко встречающийся ЛД в тексте документа, и частотой встречаемости ЛД w, поскольку очевидно, что:

$$lTF(w,t) = \frac{\log_2 \frac{S(L_t)}{1} - \log_2 \frac{S(L_t)}{1+k(w,L_t)}}{\log_2 S(L_t)}. \qquad (28)$$

Введённая величина *lTF* может быть использована вместо *TF* во всех формулах, рассмотренных в настоящей статье.

Преимуществом *lTF* перед *TF* является меньшая чувствительность к наличию текстов, сильно различающихся по длине, и связанному с этим искажению значений признаков классификации.

Следующий раздел настоящей статьи посвящён экспериментальной проверке некоторых методов классификации, рассмотренных в предыдущих разделах.

4. Описание экспериментов и их результатов

В качестве исходных данных для классификации использованы текстовые документы на русском языке (рекламные объявления), содержащие заголовок и основной текст. В исследовании рассматривалась классификация по заголовкам без учёта основного текста, поскольку при применении этого подхода были получены наилучшие результаты в рамках предыдущего исследования авторов статьи, представленного в работе [4]. Таким образом, классификация документов проводилась на основе текстового содержания их названий, а сами классифицируемые объекты представлялись векторами слов, входящих в названия документов. Средняя длина такого представления документа в словах составила 11 слов.

Для лингвистического анализа текста использован анализатор Pymorphy [10]. В проведённых экспериментах в качестве ЛД учитывались только простые слова – многословные ЛД не рассматривались, поскольку синтаксические связи между словами в тексте не выделялись.

Общий размер коллекции составил 1635 документов, количество тематических рубрик – 5. Разбиение документов по рубрикам следующее:

- магазины/товары – 995 документов;
- обучение/курсы – 262 документа;
- автомобили/мойки – 145 документов;
- активный отдых – 142 документа;
- театры/экскурсии – 91 документ.

Задача решается в закрытой однозначной постановке: каждому объявлению соответствует ровно одна тематическая рубрика.

Для классификации данные разделялись на обучающую и тестовую выборки в следующем соотношении:

- обучающая выборка – 817 документов;
- тестовая выборка – 818 документов.

В ходе эксперимента проводилась серия тестов, в каждом из которых разбиение на обучающую и тестовую выборку было случайным (равномерным) для каждой из рубрик.

Качество классификации измерялось по метрике precision (точность классификации) с микроусреднением [1].

Для обеспечения стабильности оценок качества классификации была проведена серия экспериментов (10 тестов) для каждого метода классификации. В качестве результирующих данных по каждому методу в таблицах 1 и 2 представлены следующие величины, выраженные в %:

- минимальная точность классификации;
- средняя точность классификации на серии экспериментов (медианное значение);
- максимальная точность классификации.

В Табл. 1 приведены данные по качеству классификаторов на обучающем множестве, в Табл. 2 – данные по качеству классификаторов на тестовой выборке, не вошедшей в обучающее множество.

Табл.1. Точность классификации на обучающей выборке

Метрика	Минимальная точность	Средняя точность	Максимальная точность
Euclidean	97.18482	97.931456	98.65361
Euclidean + SVM	*96.5728*	*97.58872*	*98.2864*
Cosine	95.34884	96.08323	97.30722
Cosine + SVM	*97.552*	*98.164*	*98.776*
Hellinger	94.24725	95.250918	96.32803
Hellinger + SVM	*98.6536*	*99.22888*	*99.6328*
Multinomial	91.67687	93.500613	95.22644
Multinomial + SVM	*98.164*	*98.67808*	*99.2656*
Euclidean + Cosine + Hellinger + Multinomial + SVM	*98.5312*	*99.06976*	*99.6328*
XT3	98.40881	98.886169	99.5104
XT3 + SVM	*99.0208*	*99.40024*	*99.7552*
XT3.1	99.14321	99.449203	**99.7552**
XT3.1 + SVM	*99.2656*	*99.47368*	*99.6328*

Табл.2. Точность классификации на тестовой выборке

Метрика	Минимальная точность	Средняя точность	Максимальная точность
Euclidean	86.43032	88.31296	89.8533
Euclidean + SVM	*83.1296*	*85.7335*	*88.3863*
Cosine	85.69682	87.102689	88.26406
Cosine + SVM	*85.2078*	*88.38631*	*90.3423*
Hellinger	78.97311	80.929098	82.76284
Hellinger + SVM	*86.3081*	*88.44744*	*91.687*
Multinomial	53.05623	55.403423	58.92421
Multinomial + SVM	*79.9511*	*82.32273*	*84.8411*
Euclidean + Cosine + Hellinger + Multinomial + SVM	*86.5526*	*87.78731*	*89.6088*
XT3	84.59658	87.090465	88.87531
XT3 + SVM	*87.4083*	*89.31541*	*91.3203*
XT3.1	87.04156	88.202933	89.48655
XT3.1 + SVM	*86.9193*	*87.71395*	*88.7531*

Результаты метода классификации на основе XT3 (формула (21)) обозначены аббревиатурой «XT3». Результаты метода классификации на основе модифицированной XT3 (формула (26), вместо оценки *TF* использована оценка *lTF*) обозначены аббревиатурой «XT3.1». Остальные методы обозначены в соответствии с используемыми метриками (формулы (8) – (11)).

Результаты экспериментов показывают следующее.

1. Все методы классификации, проверенные в ходе экспериментов, демонстрируют высокую точность на обучающей выборке: на уровне 95% и выше. Это означает, что сформированные по результатам обучения признаки классификации хорошо разделяют обучающую выборку на рубрики.

2. На тестовой выборке качество классификации снижается на ~10%.

3. Использование метода SVM даёт небольшое улучшение качества для отдельных методов (Multinomial, Hellinger). Однако его применение значительно увеличивает время работы классификатора, что требует усложнения реализации за счёт использования параллельных алгоритмов [6]. Для некоторых методов увеличение качества незначительное (XT3, Cosine) или наблюдается ухудшение (Euclidean, XT3.1)

4. Точность методов на основе XT3 и её модификации соответствует лучшим результатам других методов. Модифицированная XT3 при этом даёт наименьший разброс точности для минимального и максимального результата.

5. Высокие и достаточно близкие показатели точности классификации всех методов позволяют предположить, что на имеющемся тестовом множестве практически достигнуто максимально возможное качество для методов,

использующих отдельные слова в качестве признаков классификации.

Заключение

Предложенный в статье метод автоматической классификации коротких текстовых документов на основе ХТЗ и её модификации показал высокие результаты по точности классификации. Разработанный автоматический классификатор обладает простотой реализации и вычислительной эффективностью.

В дальнейшем планируется:

1) оценить качество классификации с помощью ХТЗ на полных текстах документов с учётом составных (двусловных) ЛД;

2) исследовать вклад формулы *lTF* в качество результатов классификации полнотекстовых документов в сравнении с формулой *TF*.

3) реализовать композицию алгоритмов машинного обучения («Boosting») [27] для оценки практического порога качества классификации на заданной обучающей выборке;

4) выполнить оценку динамики практического порога качества классификации в зависимости от размера обучающей выборки.

Важно отметить, что предложенный подход к оценке значимости ЛД ЕЯ позволяет в рамках единой информационно-аналитической системы построить модель значимости ЛД текстов и реализовать на основе этой модели комплекс алгоритмов, включающий:

– полнотекстовую классификацию;

– информационный поиск с учётом тематики;

– поиск тематически похожих документов;

– формирование ключевых ЛД документов с учётом их тематики;

– выявление сходства и различия между коллекциями текстов.

Литература

1 М. Агеев, И. Кураленок, И.Некрестьянов Официальные метрики РОМИП 2006 [Электронный ресурс] – Режим доступа: http://romip.ru/romip2006/appendix_a_metrics.pdf, свободный. Проверено 19.08.2012.

2 Андреев, А.М. Модели и методы автоматической классификации текстовых документов /А.М. Андреев, Д.В. Березкин, В.В. Сюзев, В.И. Шабанов, Вестн. МГТУ, Сер. Приборостроение. – М.: Изд-во МГТУ.– 2003.– №3.

3 Вейзе А. А. О ядерных текстах и их получении путем компрессии // Проблемы текстуальной лингвистики /Под. ред. проф. В.А. Бухбиндера. — Киев, 1983.

4 Драль А.А., Мбайкоджи Э. Классификация коротких текстовых документов, Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем: Тезисы докладов Всероссийской конференции с международным участием – М.:РУДН.- 2012.- С.121-123.

5 Завьялова О.С., Киселёв А.А., Осипов Г.С., Смирнов И.В., Тихомиров И.А. Соченков И.В. Система интеллектуального поиска и анализа информации «Exactus» на РОМИП-2010 [Электронный ресурс] – Режим доступа: http://romip.ru/romip2010/04_exactus.pdf, свободный. Проверено 18.08.2012.

6 Котельников Е. В., Пескишева Т.А. Параллельная система автоматической классификации // международный журнал Программные системы и продукты [Электронный ресурс] – Режим доступа: http://www.swsys.ru/index.php?page=article&id=3008, свободный. Проверено 19.08.2012.

7 Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. Введение в информационный поиск - 2011 г. С. 263 – 294.

8 Д.В. Ландэ, А.Т. Дармохвал, А.Ю. Морозов. Подход к выявлению дублирования сообщений в новостных информационных потоках. [Электронный ресурс] – Режим доступа: http://elib.lvk.cs.msu.su/papers/documents0/1/http:zSzzSzwww.rcdl2006.uniyar.ac.ruzSzpaperszSzpaper_71_v2.pdf, свободный. Проверено 18.08.2012.

9 Лотман Ю.М. Структура художественного текста. — М., 1970; Общение. Текст. Высказывание. — М., 1989.

10 Морфологический анализатор pymorphy [Электронный ресурс] – Режим доступа: http://pymorphy.readthedocs.org/en/v0.5.6/index.html, свободный. Проверено 19.08.2012.

11 Попов А. Поиск в Интернете – внутри и снаружи // А. Попов, Журнал "Интернет", #2(7) 1998 г. [Электронный ресурс] – Режим доступа: http://www.shipbottle.ru/projects/txt/internet_2_1998/index.shtml, свободный. Проверено 22.05.2009.

12 Севбо И. П. Структура связного текста и автоматизация реферирования. — М., 1969.

13 Стратонович Р. Л. Теория информации // Р. Л. Стратонович, М.: Сов. Радио, 1975, 424 с. [Электронный ресурс] – Режим доступа: http://www.polytech.poltava.ua/lib/resurs/tik/stratonovich.pdf, свободный. Проверено 29.05.2009.

14 Тихомиров И.А., Соченков И.В. Метод динамической контентной фильтрации сетевого трафика на основе анализа текстов на естественном языке, Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2008. Т. 6. № 2. С. 94-100.

15 «Яндекс подсчитал русскоязычных пользователей твиттера». // Lenta.ru [Электронный ресурс] – Режим

доступа: http://lenta.ru/news/2011/08/04/rustwitter/, свободный. Проверено 18.08.2012.
16 Amati, G. Probabilistic models of information retrieval based on measuring the divergence from randomness / G. Amati and C. J. Van Rijsbergen, The Information Retrieval Group, 20(4):357-389, 2002. [Электронный ресурс] – Режим доступа: http://ir.dcs.gla.ac.uk/terrier/publications/p357-amati.pdf, свободный. Проверено 22.05.2009.
17 Department of Statistics, Stanford University, Fall, 2008// Correspondence Analysis and Related Methods. [Электронный ресурс] – Режим доступа: http://www.econ.upf.edu/~michael/stanford/maeb4.pdf, свободный. Проверено 18.08.2012.
18 Han, E. Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification / E. Han, G. Karypis, V. Kumar, 16th International Conference on Machine Learning – Denver, 1999. – P.p. 41-56.
19 Information Gain. Universitatea Tehnica din Cluj-Napoca [Электронный ресурс] – Режим доступа: http://ftp.utcluj.ro/pub/users/nedevschi/AV/12_FeatureSelectionPerformanceEvaluation/entropyEtc.pdf, свободный. Проверено 18.08.2012.
20 Koller, D. Hierarchically classyffying documents using very few words // Koller D., Sahami M., Proc. ICML-97. – Nashvilee, 1997 – C.170-176.
21 Lee C-H. LEARNING INDUCTIVE RULES USING HELLINGER MEASURE // Applied Artificial Intelligence , Volume 13, Number 8, 1 December 1999 , P.p. 743-762(20).
22 Masable.com//How big is the web and how fast it is growing [Электронный ресурс] – Режим доступа: http://mashable.com/2011/06/19/how-many-websites/#17197How-Fast-Is-the-Web-Growing, свободный. Проверено 18.08.2012.
23 David Meyer. Support Vector Machines.The Interface to libsvm in package e1071, Technische Universität Wien, Austria, 2011 [Электронный ресурс] – Режим доступа: http://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf, свободный. Проверено 18.08.2012.
24 Netcraft// August 2012 Web Server Survey. [Электронный ресурс] – Режим доступа: http://news.netcraft.com/archives/2012/08/02/august-2012-web-server-survey.html, свободный. Проверено 18.08.2012.
25 Robertson, S. E. Probabilistic models of indexing and searching. In R.N. Oddy // S.E. Robertson, C.J. van Rijsbergen, P.W. Williams, Information Retrieval Research, pages 35-56, London, 1981. Butterworths. [Электронный ресурс] – Режим доступа: http://www.soi.city.ac.uk/~ser/papers/Robertson_vanRijsbergen_Porter.pdf, свободный. Проверено 22.05.2009.
26 The BM25 Weighting Scheme // Xapian Open Source Search Engine Library. / [Электронный ресурс] – Режим доступа: http://xapian.org/docs/bm25.html, свободный. Проверено 22.05.2009
27 Yoav Freund and Robert E. Schapire A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, Journal of Computer and System Sciences, 1997, 55(1): P.p. 119-139.

Мбайкоджи Эмели. Магистрант Российского университета дружбы народов. В 2010 окончила бакалавриат факультета физико-математических и естественных наук РУДН. Автор трех печатных работ. Область научных интересов: машинный анализ естественно-языковых текстов, компьютерная лингвистика, методы классификации, поиска и анализа текстовой информации. E-mail: Emily.rudn@gmail.com

Драль Алексей Александрович. Аспират МГУ им. Ломоносова. В 2010 году окончил механико-математический факультет МГУ. Автор двух печатных работ. Область научных интересов: машинный анализ естественно-языковых текстов, компьютерная лингвистика, методы извлечения информации из текстов. E-mail: aadral@gmail.com

Соченков Илья Владимирович. Инженер-исследователь лаборатории интеллектуальных динамических систем ИСА РАН, ведущий программист ООО "Технологии системного анализа ". Окончил Российский университет дружбы народов в 2009 году. Автор 20 научных работ. Область научных интересов: интеллектуальные методы поиска и анализа информации, обработка больших массивов данных, защита компьютерных сетей, контентная фильтрация, компьютерная лингвистика. E-mail: sochenkov@isa.ru