ISBN 5-94052-066-0 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ. ПЕРЕСЛАВЛЬ-ЗАЛЕССКИЙ, 2004

УДК 519.767.6

М. Бабенко, Е. Куршев, О. Одинцов, Е. Сулейманова,
А. Чеповский

Система классификации текстов информационных сообщений на русском языке «АКТИС»

Аннотация. В статье описывается структура системы классификации текстовых сообщений, алгоритмы обучения и классификации, используемые в системе. Приводятся результаты тестирования системы, значения коэффициентов точности и полноты в зависимости от различных наборов признаков, используемых при классификации.

Ключевые слова и фразы: классификация текстов, компьютерная лингвистика.

Введение

В современном мире, характеризуемом быстрорастущими объемами информации, одним из основных средств обработки и организации текстовых данных становится классификация текстов. Классификация текстов используется для обработки информационных сообщений, для поиска необходимой информации в больших объемах текстов, например, в сети Internet.

Цель классификации текстов — разделение документов на фиксированное число предопределенных категорий, или классов. Каждый документ может попасть во много, только одну, либо вообще ни в одну из категорий. При использовании автоматического обучения цель — обучить классификаторы на примерах, которые соответствуют данным категориям.

Задачу классификации текстов можно интерпретировать по-разному. С математической точки зрения — это задача распознавания

Работа выполнена в рамках суперкомпьютерной программы «СКИФ» Союзного государства.

образов в алгебраической постановке, а значит, для её решения можно использовать те же принципы, что и для моделирования поведения живых организмов, оценки технологических процессов, распознавания графических объектов и пр. При таком подходе для каждого объекта выделяются вектора признаков. В случае текстов признаками являются слова и взаимосвязанные наборы слов, содержащиеся в тексте. В результате обучения информация о соответствии признаков классам текстов сводится в информационную матрицу. Каждый элемент информационной матрицы определяет вес признака при принятии решения о принадлежности признака к данному классу.

В настоящее время ведутся активные разработки в этом направлении, и ожидается, что системы классификации, построенные по таким принципам, смогут решать задачи определения эмоционально-психологической окраски текстов, выделения системных установок, заложенных в тексте, и др.

1. Структура системы

Одной из таких систем является система классификации текстов информационных сообщений «АКТИС», разработаннная и реализованная в рамках программы «СКИФ» Союзного государства. Система при своей работе использует следующие модули:

(1) Модуль предварительной обработки текста. Осуществляет переход между таблицами символов текста и фильтрацию элементов форматирования исходного документа. На вход модуля подается поток символов текстового документа в исходном формате. На выход модуля поступает отфильтрованный поток символов, содержащий минимум элементов форматирования, в кодировке Windows-1251.

(2) Модуль графематического разбора. Осуществляет выделение слов (точнее, так называемых токенов — словоформ или спецсимволов) из состава отфильтрованного текста. Все токены приводятся к единому нижнему регистру, для каждого заводится набор признаков, описывающих в обобщенном виде его исходное представление в тексте (слово начиналось с большой буквы, набрано большими буквами, представляет собой спецсимвол, и проч.). На вход модуля поступает отфильтрованный текст, выходом является последовательность токенов, снабженных дополнительными (графематическими) признаками.

(3) Модуль морфологического разбора. Определяет морфологические характеристики словоформы и приводит ее к каноническому виду. Входные словоформы преобразуются в каноническое представление и снабжаются набором морфологических признаков (каноническая форма и приписанный ей вектор значений грамматических признаков далее называется морфологическим элементом).
(4) Модуль постморфологического разбора. Осуществляет группировку морфологических элементов на основе некоторого набора правил (анализ устойчивых словосочетаний, слов, написанных через дефис, составных фамилий и имен). Данный этап необходим для упрощения формирования именных групп.
(5) Модуль неполного синтаксического разбора. Для построения именных групп используется упрощенная синтаксическая модель. На вход модуля подается набор морфологических элементов, полученных в результате постморфологического разбора и образующих одно предложение исходного текста. Выходом этого модуля являются предложные и беспредложные именные группы. Набор результатов работы этого модуля для конкретного текста образует множество признаков данного текста, которые могут быть использованы для классификации.
(6) Модуль обучения. Составляет список признаков на основе обработки обучающих текстов. Формирует матрицу значимости признаков для того или иного класса.
(7) Модуль классификации. На основе информации, полученной при обучении системы, для заданного набора признаков конкретного текста определяет степень принадлежности этого текста к классам, на множестве которых была обучена система. При попадании степени принадлежности текста некоторому классу в заданный пользователем интервал считается, что текст принадлежит этому классу.

В реализуемой модели каждое слово русского языка относится к одной из 22 групп, которые называются морфологическими категориями. Понятие морфологической категориии близко к понятию части речи в грамматике, хотя в общем случае не совпадает с ним. Каждая морфологическая категория характеризуется определенным

вектором грамматических категорий (например, для существительного — род, одушевленность, число, падеж) и своим вектором словоформ.

Рассматриваются следующие морфологические категории: (номер соответствует признаку в словаре и результатах разбора)

- Неизменяемое слово,
- Существительное,
- Прилагательное,
- Глагол несовершенного вида,
- Предлог,
- Глагол совершенного вида,
- Количественное числительное,
- Порядковое числительное,
- Местоимение,
- Местоименное прилагательное,
- Собирательное числительное,
- Сокращение,
- Латинское слово,
- Аббревиатура,
- Фамилия,
- Имя,
- Отчество,
- Причастие,
- Союз,
- Наречие,
- Частица,
- Междометие,
- Топоним,
- Субстантивированное прилагательное.

Как уже было сказано, задача модуля синтаксического анализа предложения состоит в выделении в тексте именных групп, беспредложных (ИГ) и предложных (ПГ). Алгоритм синтаксического анализа опирается на принципы, описанные в [1]. Синтаксический анализ состоит из двух основных этапов:

(1) Установление между словами предложения бинарных подчинительных связей различного типа:
   - беспредложное примыкание существительных и их аналогов,

- адъективное согласование (т. е. согласование существительных и их аналогов с полными прилагательными, причастиями, порядковыми числительными),
- адвербиальное примыкание (наречия к прилагательному),
- управление (предлога существительными и их аналогами),
- связь в количественных конструкциях.

(2) На основе установленных синтаксических связей строятся именные группы различной длины.

Единицей входных данных для модуля синтаксического анализа является предложение. Предложение представлено в виде вектора результатов морфологического анализа для каждого слова предложения. Результатом работы модуля синтаксического анализа является упорядоченный список именных групп, содержащихся в разбираемом предложении.

2. Алгоритмы обучения и классификации

В результате обработки множества документов обучающей выборки, относящихся к некоторому классу C (релевантных классу), выделяется множество признаков, так или иначе характеризующих этот класс. Такими признаками могут быть как построенные именные группы, так и слова исходного текста, относящиеся к определенным пользователем морфологическим категориям. Каждому признаку t относительно некоторого класса C ставится в соответствие набор следующих статистических параметров:.

- $R^+(t, C)$ — количество документов, релевантных классу C, содержащих признак t.
- $N^+(t, C)$ — количество документов, не релевантных классу C, содержащих признак t.
- $R^-(t, C)$ — количество документов, релевантных классу C, содержащих признак t.
- $N^-(t, C)$ — количество документов, не релевантных классу C, и не содержащих признак t.

Указанные параметры используются для оценки следующих вероятностных характеристик признака t относительно класса C:

$P(C|t) = \frac{R^+}{R^+ + N^+}$ — условная вероятность того, что текст, содержащий термин t, принадлежит классу C;

$P(\bar{C}|t) = 1 - P(C|t) = \frac{N^+}{R^+ + N^+}$ — условная вероятность того, что текст, содержащий термин t, не принадлежит классу C;

$P(C|\bar{t}) = \frac{R^-}{R^- + N^-}$ — условная вероятность того, что текст, не содержащий термин t, принадлежит классу C;

$P(\bar{C}|\bar{t}) = \frac{N^-}{R^+ + N^+}$ — условная вероятность того, что текст, не содержащий термин t, не принадлежит классу C.

Считается, что признак t тем более характерен для класса C, чем выше вероятности $P(C|t)$ и $P(\bar{C}|\bar{t})$ и чем меньше вероятности $P(C|\bar{t})$ и $P(\bar{C}|t)$. На основании данного утверждения можно построить следующий критерий релевантности:

$$\rho = \frac{P(C|t)P(\bar{C}|\bar{t})}{P(\bar{C}|t)P(C|\bar{t})} \tag{1}$$

На основании собранной в процессе обучения информации строится информационная матрица I размера $N \times K$, где N — количество попарно различных признаков, выделенных из всех обучающих текстов, K — количество классов, на которых было проведено обучение.

Каждый элемент матрицы I_{tC} представляет собой упорядоченную пару характеристик $\langle \chi^2, \rho \rangle$, где χ^2 — коэффициент контингенции, а ρ — коэффициент релевантности. Для обоих коэффициентов справедливо утверждение, что большие их значения соответствуют признакам, наиболее точно характеризующим данный класс.

Коэффициент контингенции определяется по следующему соотношению [2]:

$$\chi^2_{i,j} = \frac{M \cdot \left(R^+_{i,j} \cdot N^-_{i,j} - R^-_{i,j} \cdot N^+_{i,j}\right)^2}{\left(R^+_{i,j} + R^-_{i,j}\right) \cdot \left(N^+_{i,j} + N^-_{i,j}\right) \cdot \left(R^+_{i,j} + N^+_{i,j}\right) \cdot \left(R^-_{i,j} + N^-_{i,j}\right)}, \tag{2}$$

где M — общее количество документов в обучающей выборке, i — номер признака, j — номер класса, а остальные параметры определены выше.

Коэффициент релевантности определяется по формуле (1) после подстановки значений условной вероятности и упрощения выражения:

$$\rho_{i,j} = \frac{R^+_{i,j}/R^-_{i,j}}{N^+_{i,j}/N^-_{i,j}}, \tag{3}$$

где все параметры аналогичны параметрам формулы (2).

После создания описания классов (обучения) возможна классификация текстов. При классификации текста выполняется алгоритм морфологического анализа, выделяющий множество терминов текста, и выполняется алгоритм классификации, определяющий меру принадлежности текста к каждому из классов.

Исходными данными для алгоритма статистического обучения служит множество классов $C = \{C_1 \dots C_K\}$, каждый из которых содержит идентификатор класса и множество текстовых документов, образующих этот класс. Текстовый документ (в рамках данного алгоритма), в свою очередь, представляется множеством признаков, выделенных из него с помощью алгоритма построения именных групп или непосредственным отбором слов исходного текста, прошедших этап морфологического разбора.

Значения коэффициентов контингенции и релевантности определяются согласно соотношениям, приведенным выше.

Необходимо заметить, что в целях оптимизации работы алгоритма полезными являются следующие дополнительные соотношения:

- $R^-_{j,i} = R_j - R^+_{j,i}$, где R_j — количество текстов, принадлежащих (релевантных) классу C_j;
- $N^-_{j,i} = N_j - N^+_{j,i}$, где N_j — количество текстов, не релевантных классу C_j.

Формальное определение алгоритма обучения с учетом вышеприведенных соотношений приведено на рисунке 1.

$\forall C_j \in C$:
  $\forall t_i \in T$:
    $\rho_{i,j} = \frac{R^+_{i,j}/R^-_{i,j}}{N^+_{i,j}/N^-_{i,j}}$
    $\chi^2_{i,j} = \frac{M \cdot \left(R^+_{i,j} \cdot N^-_{i,j} - R^-_{i,j} \cdot N^+_{i,j}\right)^2}{\left(R^+_{i,j}+R^-_{i,j}\right) \cdot \left(N^+_{i,j}+N^-_{i,j}\right) \cdot \left(R^+_{i,j}+N^+_{i,j}\right) \cdot \left(R^-_{i,j}+N^-_{i,j}\right)}$
    $I_{i,j} = \left\langle \chi^2_{i,j}, \rho_{i,j} \right\rangle$
return I
end

Рис. 1. Алгоритм обучения

Входными данными для алгоритма классификации являются информационная матрица, полученная на этапе обучения системы, а также набор управляющих коэффициентов, описанных ниже.

Результатом работы является множество классов, к которым, по мнению системы, может принадлежать данный текстовый документ.

Для каждого признака определяется количество его вхождений в классифицируемый текст.

Так же, как и при обучении, вычисляются два коэффициента, определяющих меру типичности текста для каждого из классов, — коэффициент контингенции и коэффициент релевантности. Среди каждого из множеств этих коэффициентов определяется максимальное значение контингенции и релевантности M_k и M_r. На основании полученных данных, а также управляющих коэффициентов, заданных пользователем, определяются два интервала включения, управляющие результатами работы алгоритма.

Коэффициент контингенции текста относительно класса C_i определяется так:

$$\chi_j^2 = \sum_i \left(\chi_{i,j}^2 \cdot count(v_i)\right),$$

где $count(v_i)$ — количество вхождений i-го признака в рассматриваемый текст. По сути, данное выражение представляет собой скалярное произведения вектора признаков документа и соответствующего классу C_i столбца информационной матрицы.

Коэффициент релевантности текста классу C_j:

$$\rho_j = \sum_i \left(\rho_{i,j} \cdot count(v_i)\right),$$

где все параметры определяются аналогично предыдущей формуле.

Управляющими параметрами для данного алгоритма являются:

- m_k и m_r — минимальные значения коэффициентов контингенции и релевантности соответственно;
- D_k и D_r — относительные значения отклонения коэффициентов контингенции и релевантности от их максимальных значений.

Интервал включения для контингенции:

$$S_k = [\max\{m_k, M_k(1 - D_k)\}; \infty],$$

для релевантности:

$$S_r = [\max\{m_r, M_r(1 - D_r)\}; \infty].$$

При одновременном попадании коэффициента контингенции текста для i-го класса в S_k, а коэффициента релевантности текста относительно i-го класса в S_r считается, что рассматриваемый текст релевантен классу C_i.

Алгоритм классификации представлен на Рис. 2.

$$
\begin{aligned}
&\forall C_j \in C : \\
&\quad \chi_j^2 = \sum_i \left(\chi_{i,j}^2 \cdot count(v_i)\right) \\
&\quad \rho_j = \sum_i \left(\rho_{i,j} \cdot count(v_i)\right) \\
&M_k = \max_j \left\{\chi_j^2\right\} \\
&M_r = \max_j \left\{\rho_j\right\} \\
&S_k = [\max\{m_k, M_k(1 - D_k)\}; \infty] \\
&S_r = [\max\{m_r, M_r(1 - D_r)\}; \infty] \\
&R = \emptyset \\
&\forall C_j \in C : \\
&\quad \text{if } \chi_j^2 \in S_k \wedge \rho_j \in S_r : \\
&\qquad R = R \cup \{C_j\} \\
&\text{return R} \\
&\text{end}
\end{aligned}
$$

Рис. 2. Алгоритм классификации

3. Результаты тестирования

Для исследования влияния различных лингвистических характеристик на качество классификации текстовых сообщений был проведен ряд тестовых запусков системы с разными значениями множества возможных типов признаков, используемых при классификации. Исследовалась работа системы в следующих режимах:

(1) Только беспредложные именные группы. В состав вектора признаков, используемых при классификации, входят только беспредложные именные группы. Остальные признаки текста игнорируются.

(2) Именные группы, включая предложные группы.
(3) Только имена существительные.
(4) Именные группы и существительные.
(5) Имена существительные и глаголы.
(6) Все возможные признаки, исключая именные группы.
(7) Все возможные признаки, включая именные группы.

Для обучения системы использовался тестовый набор текстов, состоящий из 3-х классов (условные названия «36», «83», «89»), и 2522 текстовых документов общим объемом 4288014 байт. Классификация осуществлялась на наборе из 300 документов общим объемом 432978 байт, предварительно разбитых по классам для оценки качества классификации. Для количественной оценки качества классификации использовались следующие соотношения:

Полнота классификации в классе C_i:

$$X_f^i = \frac{N_{sr}^i}{N_s^i},$$

где N_{sr}^i — количество текстовых сообщений, корректно отнесенных системой к классу C_i, N_s^i — количество текстовых сообщений, априорно отнесенных к классу C_i.

Полнота классификации в классе C_i:

$$X_{acc}^i = \frac{N_{sr}^i}{N_r^i},$$

где N_{sr}^i определяется аналогично предыдущей формуле, а N_r^i обозначает количество текстовых сообщений, отнесенных к классу C_i по результатам классификации.

Кроме параметров качества для каждого из классов, также оцениваются суммарные значения средней точности и полноты, определяемые следующим образом:

$$X_f^\Sigma = \frac{\sum\limits_i N_{sr}^i}{\sum\limits_i N_s^i}$$

— суммарная средняя полнота классификации;

$$X_{acc}^\Sigma = \frac{\sum\limits_i N_{sr}^i}{\sum\limits_i N_r^i}$$

— суммарная средняя точность классификации.

При тестировании системы также варьировались и параметры классификации (минимальное значение и порог контингенции и релевантности).

Результаты тестирования влияния лингвистических характеристик на качество классификации приведены в таблицах 1, 2, 3.

В качестве результатов тестов приводятся как средние значения качества, так и значения качества, полученные по отношению к двум фиксированным классам (83 и 36), позволяющие оценить пиковые значения точности и полноты классификации.

Параметры	Средние значения		min значения		Класс «83»		Класс «36»	
	R	P	R	P	R	P	R	P
Только ИГ	88%	97%	81%	97%	81%	98%	96%	97%
ИГ + ПГ	90%	97%	83%	97%	83%	96%	98%	98%
Только сущ.	85%	95%	76%	87%	83%	100%	98%	87%
ИГ + сущ.	92%	99%	88%	96%	90%	100%	98%	96%
Сущ. + глаг.	85%	93%	73%	84%	83%	100%	98%	84%
Все без ИГ	83%	93%	70%	84%	86%	100%	92%	84%
Все с ИГ	91%	97%	84%	92%	93%	100%	97%	92%
Значения параметров классификации: min контингенции 0.01, порог контингенции 0.3, min релевантности 0.01, порог релевантности 0.3. R — полнота, P — точность.								

Таблица 1. Качество работы алгоритма классификации (полнота и точность) для первого набора параметров

Наглядное представление этих результатов приведено на Рис. 3 и Рис. 4 (порог классификации для критериев контингенции и релевантности устанавливается одинаковым).

Параметры	Средние значения		min значения		Класс «83»		Класс «36»	
	R	P	R	P	R	P	R	P
Только ИГ	92%	94%	89%	96%	89%	96%	96%	96%
ИГ + ПГ	93%	94%	90%	92%	90%	92%	98%	95%
Только сущ.	90%	91%	82%	84%	91%	95%	98%	84%
ИГ + сущ.	96%	92%	91%	90%	97%	91%	99%	90%
Сущ. + глаг.	90%	89%	81%	82%	92%	92%	98%	82%
Все без ИГ	86%	85%	72%	73%	93%	93%	93%	73%
Все с ИГ	95%	90%	89%	83%	99%	93%	97%	83%
Значения параметров классификации: min контингенции 0.01, порог контингенции 0.5, min релевантности 0.01, порог релевантности 0.5. R — полнота, P — точность.								

Таблица 2. Качество работы алгоритма классификации (полнота и точность) для второго набора параметров

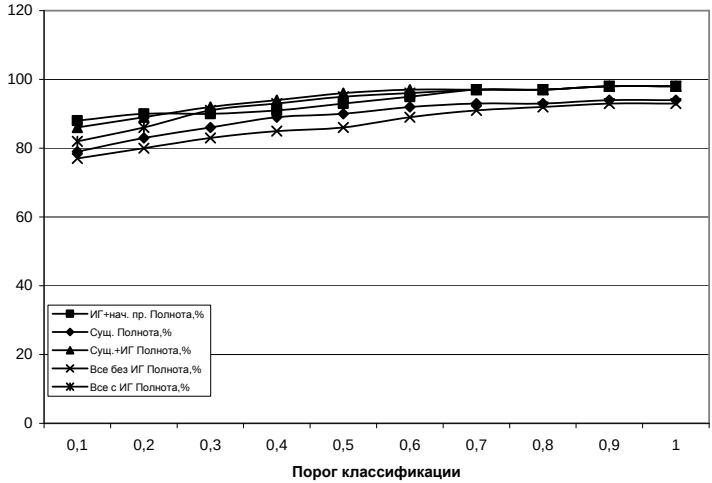

Рис. 3. График влияния лингвистических парамет-

Параметры	Средние значения		min значения		Класс «83»		Класс «36»	
	R	P	R	P	R	P	R	P
Только ИГ	95%	85%	92%	83%	92%	89%	98%	84%
ИГ + ПГ	97%	85%	95%	84%	96%	85%	99%	84%
Только сущ.	93%	82%	87%	78%	93%	85%	99%	78%
ИГ + сущ.	97%	85%	95%	82%	98%	82%	99%	83%
Сущ. + глаг.	92%	81%	86%	82%	92%	82%	99%	76%
Все без ИГ	86%	85%	72%	73%	93%	93%	93%	73%
Все с ИГ	95%	90%	89%	83%	99%	93%	97%	83%
Значения параметров классификации: min контингенции 0.01, порог контингенции 0.7, min релевантности 0.01, порог релевантности 0.7. R — полнота, P — точность.								

Таблица 3. Качество работы алгоритма классификации (полнота и точность) для третьего набора параметров

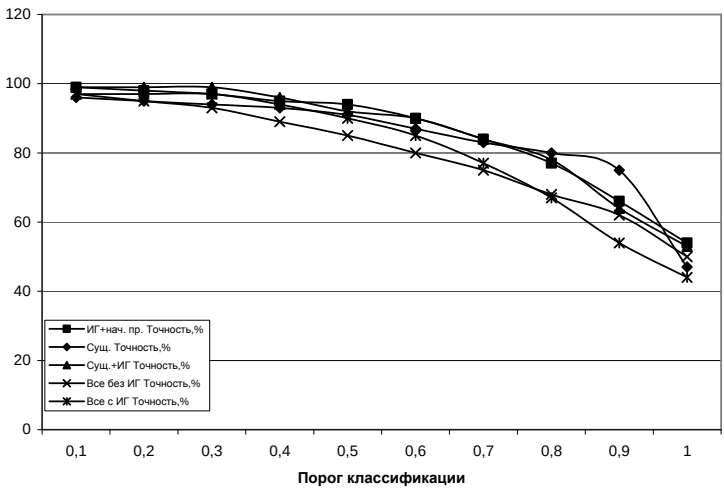

Рис. 4. График влияния лингвистических парамет-

Анализируя полученные данные, можно сделать вывод, что использование именных групп в подавляющем большинстве случаев дает положительный эффект по отношению как к точности классификации, так и к ее полноте. Из остальных категорий признаков наибольшее влияние на качество классификации оказывает использование имен существительных. Использование всех доступных морфологических категорий в качестве признаков часто отрицательно сказывается на полноте и точности классификации. Это может свидетельствовать о низкой информативности отдельных морфологических категорий (таких, например, как глаголы) при использовании их в качестве самостоятельных признаков в задаче классификации.

Благодарности

Авторы благодарны сотрудникам ИЦМС ИПС РАН за помощь в подготовке пакета «АКТИС» к прохождению госиспытаний, особые благодарности Коваленко М. Р. за реализацию дистрибутивного пакета программ «АКТИС» и за реализацию корректной работы русификации интерфейса пакета в ОС Linux.

Список литературы

[1] Белоногов Г. Г., Новоселов А. П. Автоматизация процессов накопления, поиска и обобщения информации. — М.: Наука, 1979. ↑1

[2] Edwards P., Bayer D., Green C. L., Payne T. R. *Experience with Learning Agents which Manage Internet-Based Information* // AAAI 1996 Stanford Spring Symposium on Machine Learning in Information Access ред. Hearst M. A., Hirsh H.: AAAI, 1996, с. 31–40. ↑2

Московский государственный университет

Исследовательский центр искусственного интеллекта ИПС РАН

Московский государственный технический университет

Исследовательский центр медицинской информатики ИПС РАН

M. Babenko, E. Kourshev, O. Odintsov, E. Suleimanova, A. Chepovsky. *AKTIS – a System for Russian Language Text Categorization .* (in russian.)

Abstract. The paper presents a system for textual message categorization. The system structure, as well as the training and categorization algorithms are outlined. An overview of evaluation results is given, including recall and precision values for different sets of categorization attributes.