ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2016, № 1, с. 115–123

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

УДК 004.93'12

АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТОВ НА ОСНОВЕ ГЕНЕТИЧЕСКОГО АЛГОРИТМА И СОВМЕСТНОЙ КЛАСТЕРИЗАЦИИ СЛОВ И ДОКУМЕНТОВ*

© 2016 г. Е. В. Котельников, М. В. Плетнева

Киров, Вятский государственный гуманитарный ун-т
e-mail: Kotelnikov.ev@gmail.com

Поступила в редакцию 21.04.14 г., после доработки 24.06.15 г.

Предлагается новый метод анализа тональности текстов, основанный на вычислении весов оценочных слов, позволяющий автоматически распознавать позитивную или негативную тональность, выраженную в тексте по отношению к некоторому объекту. Проблема определения весов оценочных слов рассматривается как оптимизационная задача, критерием которой является максимизация выбранной метрики качества анализа тональности. Для сокращения пространства поиска оптимальных весов оценочных слов в методе используется совместная кластеризация, в результате которой выделяются сильно связанные группы оценочных слов и текстовых документов. Веса оптимизируются на основе генетического алгоритма независимо для каждого кластера. Эксперименты на текстовых коллекциях семинара РОМИП подтверждают эффективность предложенного метода. Областью практического применения метода является компьютерная поддержка разнообразных исследований, включающих анализ мнений, – социологических, политологических, маркетинговых и др.

DOI: 10.7868/S0002338815060104

Введение. В настоящее время одним из передовых направлений в компьютерной лингвистике является анализ тональности текстов (text sentiment analysis) или по-другому автоматическая классификация текстов по тональности. *Тональность текста* – это эмоциональная оценка, выраженная в тексте по отношению к некоторому объекту. Тональность представляется определенной шкалой, имеющей не менее двух значений. Наиболее распространенными шкалами выступают двухзначная (позитивная тональность – негативная тональность) и трехзначная (добавляется нейтральная тональность).

Интенсивное развитие этой области в последние 10–15 лет связано, во-первых, с повсеместным распространением Интернета, что привело к возможности публикации мнений множества людей; во-вторых, с появлением мощных инструментов машинного обучения, таких, как метод опорных векторов и метод максимизации энтропии, и, в-третьих, с наличием широкого диапазона потенциальных областей применения: социологические, политологические и маркетинговые исследования, рекомендательные и поисковые системы, человеко-машинный интерфейс, оценка тональности новостей и др. Подробный обзор данного направления исследований приведен в [1, 2].

Исследования в области анализа тональности ведутся в основном на материале английского языка. Для русского языка активные исследования начались только с 2012 г. на конференции по компьютерной лингвистике “Диалог” и Российском семинаре по оценке методов информационного поиска (РОМИП) [3].

В данной статье предлагается и экспериментально обосновывается метод анализа тональности текстов, в котором классификация осуществляется на основе вычисления весов оценочных слов, присутствующих в тексте. В качестве базовой выбрана двухзначная шкала тональности, но все результаты могут быть обобщены на N-значную шкалу. Для работы метода требуется словарь оценочной лексики, содержащий слова с высокой степенью эмоциональности, и текстовый корпус, включающий документы двух видов – обучающие и тестовые. Для обучающих документов известны метки, обозначающие их принадлежность к определенному значению на шкале то-

*Работа выполнена при финансовой поддержке Министерства образования и науки РФ, государственное задание ВятГГУ (код проекта 586).

нальности. Для тестовых документов такие метки также могут быть заданы, но не доступны в процессе работы метода, а используются при оценке качества классификации. Словарь оценочной лексики формируется на основе статистического анализа текстового корпуса и экспертного подхода [4].

В предлагаемом методе проблема определения весов оценочных слов рассматривается как оптимизационная задача, критерием которой является максимизация выбранной метрики качества анализа тональности. Для сокращения пространства поиска оптимальных весов в методе используется совместная кластеризация оценочных слов и текстового корпуса, в который входят одновременно обучающие и тестовые документы [5]. В результате кластеризации выделяются сильно связанные группы оценочных слов и текстовых документов. Для каждого кластера независимо вычисляются оптимальные веса оценочных слов при помощи генетического алгоритма [4], затем классифицируются тестовые документы из данного кластера с использованием найденных оптимальных весов оценочных слов. Для классификации применяется лексический метод, предложенный в [6].

Таким образом, целью статьи является экспериментальное обоснование предлагаемого метода анализа тональности с использованием текстовых коллекций отзывов о фильмах, книгах и фотокамерах семинара РОМИП 2011 г.

При совместной кластеризации информация о метках тестовых документов предполагается недоступной, что соответствует базовым принципам теории машинного обучения [7]. Следует отметить, что идея одновременной кластеризации обучающих и тестовых документов высказывалась в статье [8], но в ней не использовалась совместная кластеризация слов и документов.

1. Обзор предыдущих работ. 1.1. Подходы к взвешиванию оценочных слов. В настоящее время существует два главных подхода к автоматическому определению весов оценочных слов – на основе знаний (knowledge-based) и на основе корпусов текстов (corpus-based) [9]. В первом из этих подходов используются тезаурусы (например, WordNet), на основе которых строится граф, отражающий семантические расстояния между словами. Для определения веса слова вычисляется расстояние между ним и словами, однозначно выражающими позитивное и негативное отношение (например, “хороший” и “плохой”) [9, 10]. Во втором подходе применяются текстовые корпуса, на базе которых вычисляются различные статистические и лингвистические характеристики, позволяющие определять веса оценочных слов [11, 12]. Например, в работе [11] вычисляются три вида характеристик: частотные, на основе оценок пользователей и лингвистические, использующиеся для формирования признаковых представлений слов, которые затем классифицируются при помощи алгоритмов машинного обучения на два класса – оценочные слова и неоценочные. При этом во время классификации определяется вероятность принадлежности каждого слова к классу оценочных, которая может быть использована в качестве веса оценочного слова.

В настоящей статье предлагается способ взвешивания в рамках корпусного подхода, отличающийся от предыдущих тем, что задача определения весов оценочных слов рассматривается как проблема оптимизации, критерием которой является максимизация выбранной метрики качества анализа тональности. Однако в случае непосредственного применения техник оптимизации возникают трудности в связи с большой размерностью пространства оценочных слов. Для решения проблемы оптимизации используется генетический алгоритм, а для сокращения размерности пространства оценочных слов применяется совместная кластеризация слов и документов.

1.2. Совместная кластеризация слов и документов. Совместная кластеризация – это подход, при котором осуществляется одновременная кластеризация слов и документов. Идея совместной кластеризации строк и столбцов матрицы признаков высказывалась достаточно давно [13]. Одновременная кластеризация слов и документов была предложена в 2001 г. в двух независимых работах [5, 14].

В [5] рассматривался алгоритм, основанный на спектральном разделении двудольного графа (Bipartite Spectral Graph Partitioning). На вход алгоритма поступала матрица инцидентности “слово–документ”. Строки такой матрицы A соответствуют словам, столбцы – документам. Если i-е слово встречается в j-м документе, то элемент матрицы $A_{ij} = 1$, иначе элемент равен нулю. Преобразованная специальным образом матрица A подвергалась сингулярному разложению (Singular Value Decomposition, SVD) [15], в результате которого вычислялись вторые левые и правые сингулярные векторы. Для получения двух кластеров построенные векторы обрабатывались методом кластеризации K-средних (K-means). Формирование k кластеров осуществлялось при помощи $l = \lceil \log_2 k \rceil$ сингулярных векторов, начиная со второго. В [14] был использован такой же алгоритм на основе SVD, но для получения k кластеров предлагался менее изящный и точный ре-

курсивный способ. Совместная кластеризация применялась во многих работах по анализу тональности, например, в [16] для извлечения аспектов и связанных с ними эмоциональных оценок; в [17] для уменьшения размерности данных.

Наиболее близки разработанному в данной статье методу три работы в области анализа тональности [18–20], которые, однако, имеют важные отличия от рассматриваемого метода. В [18] так же, как и в данной статье, используется словарь оценочных слов, включаемый в матрицу "слово–документ", но не осуществляется совместная кластеризация для подбора оптимальных весов оценочных слов. В [19] аналогично предлагаемому методу в матрицу помещаются неразмеченные документы, применяются словарь и лексический метод классификации, основанный на подсчете количества позитивных и негативных слов, но в отличие от разработанного метода не учитывается коэффициент для негативных текстов (который в настоящей статье подбирается автоматически с помощью генетического алгоритма), а также не вычисляются оптимальные веса оценочных слов. Следует также отметить работу [20], в которой используется кластеризация для облегчения подбора весов. Но в предлагаемом методе применяется не обычная, а совместная кластеризация документов и слов, кроме того, в матрицу помещаются тестовые документы.

1.3. Генетический алгоритм. Генетический алгоритм – это мощный метод оптимизации функций в рамках эволюционного подхода [21, 22], который успешно применялся во многих областях, таких, как стилометрический анализ [23], распознавание рукописных символов [24], распознавание изображений [25] в основном с целью отбора наиболее значимых признаков (Feature Selection).

В области анализа тональности работ, использующих генетический алгоритм, очень немного [26, 27]. В [26] был предложен генетический алгоритм с взвешиванием на основе информационной энтропии (Entropy Weighted Genetic Algorithm, EWGA). Этот алгоритм оптимизировал процесс отбора значимых признаков, в то время как в настоящей статье генетический алгоритм применяется для взвешивания слов (Feature Weighting). Кроме того, в предлагаемом методе классификация осуществляется при помощи быстрого лексического метода, а в [26] используется метод опорных векторов, который требуется предварительно обучать. Таким образом, данный подход имеет преимущество по скорости.

В [27] распознается субъективность (Subjectivity Detection) с применением машинного обучения на базе генетического алгоритма (Genetic-Based Machine Learning, GBML). Так же как и в предыдущей работе, генетический алгоритм используется для автоматического определения наилучшего набора признаков, а не для взвешивания оценочных слов.

2. Метод анализа тональности. Метод анализа тональности на основе генетического алгоритма и совместной кластеризации слов и документов представлен на рис. 1. Метод включает три этапа: 1) вычисление весов слов входного словаря оценочной лексики, 2) совместная кластеризация слов и документов, 3) подбор оптимальных весов слов и классификация тестовых документов по кластерам.

На первом этапе вычисляются веса слов входного словаря оценочной лексики L на основе обучающей коллекции документов с использованием сглаженной обратной документной частоты (Delta Smoothed Inverse Document Frequency, DS-IDF) – метода взвешивания, который показал хорошие результаты при анализе тональности [28]:

$$w(t) = \log \frac{N_{pos} df_{neg}(t) + 0.5}{N_{neg} df_{pos}(t) + 0.5}, \tag{2.1}$$

где $w(t)$ – вес слова t; N_{pos} (N_{neg}) – общее количество позитивных (негативных) обучающих документов; $df_{pos}(t)/df_{neg}(t)$ – количество позитивных/негативных обучающих документов, которые содержат слово t. Создание входного словаря оценочной лексики L описано в разд. 3.2.

На втором этапе при помощи вызова функции CoClustering формируются k кластеров, каждый из которых содержит одновременно обучающие документы, тестовые документы и оценочные слова из словаря L (рис. 2) [5].

На третьем этапе для i-го кластера запускается функция подбора оптимальных весов слов словаря оценочной лексики L_i, входящего в i-й кластер, на основе генетического алгоритма (см. рис. 3) GeneticAlgorithm [21, 22]. В результате формируется словарь оценочных слов с оптимальными весами L_i^{opt} (рис. 1, п. 3.1). Кроме весов слов вычисляется оптимальный масштабирующий коэффициент $NegCoef_i^{opt}$, учитывающий неравномерность распределения позитивных и негативных документов. В генетическом алгоритме применяются только обучающие документы C_i^{train}, входя-

Метод SentimentAnalysis (C^{train}, C^{test}, L, k)
Входные данные:
C^{train} – коллекция обучающих документов,
C^{test} – коллекция тестовых документов,
L – словарь оценочной лексики,
k – количество кластеров.
Выходные данные:
TotalResults – список результатов классификатора на тестовых документах.

---

1. Вычисление весов слов словаря оценочной лексики L на основе метода взвешивания DS-IDF (2.1) и обучающей коллекции C^{train}.
2. Вызов функции CoClustering, которая возвращает набор из k кластеров: *Clusters* = CoClustering ($C^{train} \cup C^{test}$, L, k).
3. Для каждого кластера $\{C_i^{train}, C_i^{test}, L_i\}$ в наборе *Clusters*:
   3.1. Запуск генетического алгоритма:
   $\{L_i^{opt}, NegCoef_i^{opt}\}$ = GeneticAlgorithm (C_i^{train}, L_i).
   3.2. Добавление взвешенного словаря L к словарю L_i^{opt}:
   для каждой пары $\{word_j, weight_j\}$ из L:
   если L_i^{opt} не содержит $word_j$:
   добавление $\{word_j, weight_j\}$ в L_i^{opt}.
   3.3. Вызов функции классификации:
   Results = LexicalClassification (C_i^{test}, L_i^{opt}, $NegCoef_i^{opt}$).
   3.4. Добавление *Results* в *TotalResults*.

Рис. 1. Метод анализа тональности

Функция CoClustering (C, L, k)
Входные данные:
C – коллекция документов,
L – словарь,
k – количество кластеров.
Выходные данные:
Clusters – набор из k кластеров со словами и документами.

---

1. Формирование матрицы “слово–документ” A.
2. Формирование нормализованной матрицы: $A_n = D_1^{-1/2} \cdot A \cdot D_2^{-1/2}$,
   где $D_1 = diag(d_{11}, d_{12}, ..., d_{1i}, ...)$, $d_{1i} = \Sigma_j A_{ij}$,
   $D_2 = diag(d_{21}, d_{22}, ..., d_{2j}, ...)$, $d_{2j} = \Sigma_i A_{ij}$.
3. Вычисление $l = \lceil \log_2 k \rceil$ сингулярных векторов матрицы A_n: $u_2, ..., u_{l+1}$ и $v_2, ..., v_{l+1}$, и формирование матрицы Z:
   $$Z = \begin{bmatrix} D_1^{-1/2} U \\ D_2^{-1/2} V \end{bmatrix}, \quad U = [u_2, u_{l+1}]^{\mathrm{T}}, \quad V = [v_2, ..., v_{l+1}]^{\mathrm{T}}.$$
4. Кластеризация методом K-средних:
   Clusters = KMeans(Z, k).

Рис. 2. Функция совместной кластеризации слов и документов

щие в i-й кластер. В качестве начального приближения используются веса из словаря L. После вычисления оптимальных весов в словарь L_i^{opt} добавляются слова из исходного словаря L, которые не принадлежат кластеру, со своими весами (рис. 1, п. 3.2). Получившийся объединенный словарь поступает на вход функции LexicalClassification, реализующей лексический метод классификации (см. рис. 4) [6]. Функция возвращает результаты классификации как набор метрик качества – точности (Precision), полноты (Recall) и F1-меры (F1-measure) (рис. 1, п. 3.3) [7].

Функция GENETICALGORITHM (C, L_0, *GenCount*)
Входные данные:
C – коллекция документов,
L_0 – начальный взвешенный словарь,
GenCount – количество поколений.
Выходные данные:
L^{opt} – словарь с оптимальными весами,
$NegCoef^{opt}$ – оптимальный коэффициент для негативных документов.

1. Генерация случайных взвешенных словарей: $\{L_{1j}^{rand}\}$.
2. Формирование начальный популяции: $L_0 \cup \{L_{1j}^{rand}\}$, $\{NegCoef_{1j}\}$.
3. For $i = 1$ to *GenCount*:
   3.1. Оценка:
   для каждой пары $\{L_{ij}, NegCoef_{ij}\}$ в текущей популяции:
   $Result_{ij}$ = LEXICALCLASSIFICATION (C, L_{ij}, $NegCoef_{ij}$).
   3.2. Формирование новой популяции при помощи операторов кроссовера и мутации.
4. $\{L^{opt}, NegCoef^{opt}\} = \{L_{ij}, NegCoef_{ij}\}$ при условии: $\max\{Result_{ij}\}$.

Рис. 3. Генетический алгоритм

Функция LEXICALCLASSIFICATION (C, L, *NegCoef*)
Входные данные:
C – коллекция документов,
L – взвешенный словарь,
NegCoef – коэффициент для негативных документов.
Выходные данные:
Result – набор метрик качества: {*Precision*, *Recall*, *F*1-*measure*}.

1. Инициализация списка решений *DL*.
2. Для каждого документа d_i в коллекции C:
   Pos_i = сумма весов позитивных слов из L, присутствующих в d_i.
   Neg_i = сумма весов негативных слов из L, присутствующих в d_i.
   $Decision_i = \text{sign}(Pos_i - NegCoefNeg_i)$.
   Добавление $Decision_i$ в список решений *DL*.
3. Вычисление *Result* по списку *DL*.

Рис. 4. Функция, реализующая лексический метод классификации

Выбор количества кластеров предлагается осуществлять автоматически на основе минимального значения коэффициента вариации для распределения объектов по кластерам. Таким образом, определяется количество кластеров, при котором распределение слов и документов будет наиболее равномерным. Коэффициент вариации для k кластеров вычисляется по формуле

$$v_k = \frac{\sqrt{\frac{1}{k}\sum_{i=1}^{k}(n_i - \overline{n})^2}}{\overline{n}}, \tag{2.2}$$

где n_i – количество объектов в i-м кластере, $\overline{n}$ – среднее количество объектов в кластерах.

3. Лингвистические ресурсы. 3.1. Текстовые коллекции. Для экспериментов в работе использовались общедоступные текстовые коллекции семинара РОМИП-2011 [3]. Организаторы семинара предоставили коллекции отзывов о фильмах, книгах и фотокамерах, собранных с ре-

комендательного сервиса Imhonet (http://imhonet.ru) и с сайта Яндекс.Маркет (http://market.yandex.ru). В процессе предобработки пустые отзывы и отзывы с бессмысленным или повторяющимся содержимым были исключены, часть неразмеченных отзывов была размечена вручную, остальные удалены. Исходная десятибалльная шкала отзывов о фильмах и книгах была преобразована к бинарной по следующей схеме: $\{1\ldots5\} \to neg, \{6\ldots10\} \to pos$; исходная пятибалльная шкала для отзывов о фотокамерах также переведена в бинарную по схеме $\{1\ldots2\} \to neg$, $\{3\ldots5\} \to pos$. Затем был проведен морфологический анализ с использованием компьютерного словаря, взятого с сайта http://aot.ru, который основан на словаре А.А. Зализняка [29] и включает 174783 леммы и 3 150 188 словоформ.

После процедуры предобработки коллекция отзывов о фильмах включала 15151 отзыв, из которых 11945 (78.8%) были позитивными и 3206 (21.2%) – негативными; коллекция отзывов о книгах содержала 21707 отзывов, из которых 18385 – позитивные (84.7%), 3322 – негативные (15.3%); коллекция отзывов о фотокамерах – 10202 отзыва (9100 позитивных – 89.2%, 1102 негативных – 10.8%).

3.2. С л о в а р ь. Для генерации исходного словаря оценочной лексики L использовался следующий метод (описан на примере коллекции отзывов о фильмах) [4]. Сначала по коллекции отзывов о фильмах был сформирован словарь, содержащий около 30 000 слов. Затем на основе метода DS-IDF (2.1) вычислены веса и оставлены только 10% слов с максимальными весами. Наконец, путем просмотра были отобраны наиболее эмоционально окрашенные слова: 278 позитивных и 197 негативных. Кроме того, в словарь были включены различные виды позитивных и негативных смайликов.

Благодаря такому комбинированному экспертно-статистическому подходу в словаре оказались как предметно-независимые слова (например, *достойный*, *качественный*, *убожество*, *глупость*), так и слова, специфичные для отзывов о фильмах (например, *атмосферный*, *зрелищный*, *затянутый*, *шаблонный*).

4. Результаты экспериментов. Для повышения объективности исследования в экспериментах использовалась процедура перекрестной проверки по пяти блокам (5-fold cross-validation) [30] с сохранением пропорции позитивных и негативных отзывов. В качестве базового уровня оценки был выбран простейший классификатор, который относил все документы к наиболее частотному классу (позитивному). Для сравнения также применялся классификатор на основе метода опорных векторов (Support Vector Machine, SVM) [31], показавший хорошие результаты при классификации отзывов по тональности [32]. Для этого классификатора осуществлялся подбор параметров, в результате которого было выбрано RBF-ядро (Radial Basis Function) с параметром регуляризации $C = 10^4$. При формировании векторов для SVM-классификатора использовался описанный в разд. 3.2 словарь оценочной лексики, взвешенный с помощью метода DS-IDF (2.1), а также полный словарь обучающей коллекции, содержащий 27 300 слов, также взвешенный с помощью метода DS-IDF.

Еще одним классификатором для сравнения был выбран лексический метод анализа тональности, разработанный в [6]. В этом методе также применялся словарь, взвешенный только на базе DS-IDF.

Для предлагаемого в данной работе метода анализа тональности на основе генетического алгоритма и совместной кластеризации слов и документов использовались следующие параметры. Для генетического алгоритма: количество поколений – 200, размер популяции – 100, диапазон весов оценочных слов [0...20], разрядность представления весов – 20 бит, вероятность кроссовера – 0.9, вероятность мутации – 0.1, доля поколения, выбранная для стратегии элитизма, – 10% [22]. Совместная кластеризация была реализована при помощи математического пакета Octave (http://www.octave.org) с количеством кластеров, автоматически определяемым на основе минимального значения коэффициента вариации (2.2) для распределения объектов по кластерам.

Также проводились эксперименты по классификации отзывов только с использованием словаря данного кластера без объединения с общим словарем (исключался п. 3.2 в методе SENTIMENTANALYSIS на рис. 1). Данный вариант метода обозначен как “Предлагаемый метод со словарем кластера”. Для оценки результатов работы всех классификаторов применялись традиционные метрики – точность (P), полнота (R), F1-мера (F1), усредненные по схеме *macro* [7]. Результаты экспериментов над тестовыми коллекциями представлены в таблице.

Из таблицы видно, что, несмотря на преимущество лексического метода по полноте, предлагаемый метод опережает его (и все другие классификаторы) по точности и F1-мере. Высокая точность и относительно низкая полнота предлагаемого метода объясняются следующим образом.

Результаты классификации коллекций отзывов РОМИП

Метод	Фильмы			Книги			Фотокамеры		
	P	R	F1	P	R	F1	P	R	F1
Базовый классификатор	39.48	50.00	44.12	42.35	50.00	45.86	44.60	50.00	47.15
SVM-классификатор (словарь оценочных слов)	65.29	62.35	63.39	**68.12**	64.70	65.90	69.12	65.98	67.20
SVM-классификатор (полный словарь)	62.50	61.52	61.60	62.61	64.95	63.45	69.05	65.89	66.49
Лексический метод	63.42	**67.29**	64.26	64.17	**68.77**	65.62	66.11	**69.53**	66.93
Предлагаемый метод	**67.78**	65.02	**66.10**	67.26	65.64	**66.17**	**70.56**	67.91	**68.61**
Предлагаемый метод со словарем кластера	66.35	63.09	64.25	66.98	63.88	65.08	67.44	65.69	66.17

В схеме усреднения метрик *macro*, применяемой при сильной несбалансированности коллекции по классам, итоговые метрики вычисляются как средние значения метрик по классам, при этом все классы равнозначны независимо от количества примеров. В используемых в работе коллекциях примеры позитивного класса составляют от 78.8 до 89.2% от общего количества примеров. При подборе весов оценочных слов в генетическом алгоритме в связи со значительным преобладанием примеров позитивного класса оптимизация весов позитивных слов осуществляется эффективнее, чем негативных, что приводит к повышению точности распознавания позитивных текстов. При этом полнота для негативных текстов оказывается несколько сниженной (классификатор "предпочитает" позитивные тексты), однако относительно низкая точность для негативных текстов компенсируется высокой точностью для позитивных текстов (выше, чем в случае вычисления весов по формуле (2.1)), что приводит к преобладанию над другими методами по F1-мере.

В ходе исследования было выявлено, что в результате выполнения генетического алгоритма веса оценочных слов существенно изменяются: среднеквадратическая разность весов оценочных слов лучших популяций нулевого и последнего (200-го) поколения составляет в среднем 8.77 по всем трем коллекциям при диапазоне весов [1...20]. При этом значения F1-меры повышаются в среднем на 4.5%.

Интересно, что при классификации тестовых документов на основе только тех слов, которые входят в кластеры (вариант "Предлагаемый метод со словарем кластера"), результаты получились только незначительно хуже, чем при использовании полного словаря (разность составляет от 1.09 до 2.44%). Таким образом, одновременное применение совместной кластеризации и генетического алгоритма позволяет выявить компактный и эффективно взвешенный словарь для описания группы тестовых документов.

Также проводились эксперименты с разным количеством кластеров в предлагаемом методе (рис. 5). Максимум F1-меры достигается при количестве кластеров от 4 до 5 для разных коллекций. При этом формируется 3–4 больших кластера с большим количеством документов и слов

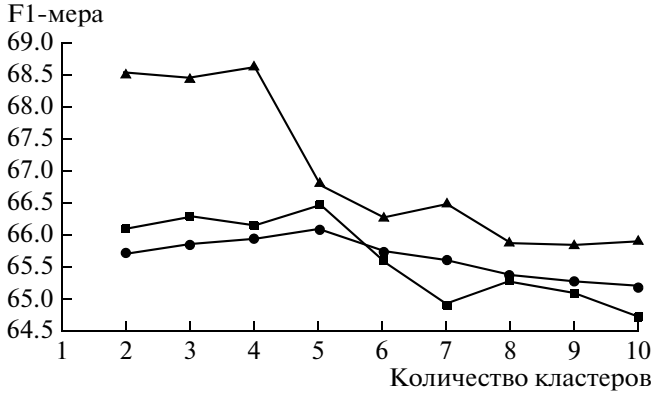

Рис. 5. Зависимость F1-меры от количества кластеров: ● – фильмы, ■ – книги, ▲ – фотокамеры

(количество слов от 48 до 230) и 1–2 малых кластера с небольшим количеством слов (количество слов от 5 до 22). Дальнейшее дробление оказывается менее эффективным. Отметим, что количество кластеров, при котором достигается минимальное значение коэффициента вариации (2.2) для распределения объектов по кластерам, совпадает с оптимальным для коллекций отзывов о фильмах ($v_{\min} = 0.487$ для 5 кластеров) и о фотокамерах ($v_{\min} = 0.319$ для 5 кластеров) и оказывается близким к нему для коллекции отзывов по книгам ($v_{\min} = 0.452$ для 4 кластеров, максимум F1-меры достигается при 5 кластерах), что позволяет сделать вывод об эффективности предложенного способа определения оптимального количества кластеров.

Заключение. В статье предложен метод анализа тональности на основе генетического алгоритма и совместной кластеризации слов и документов. В методе сначала совместно кластеризуются слова из словаря оценочной лексики и обучающие и тестовые документы, а затем вычисляются оптимальные веса оценочных слов для каждого кластера отдельно при помощи генетического алгоритма. Эксперименты на общедоступных текстовых коллекциях РОМИП подтверждают, что эффективность разработанного метода превосходит другие классификаторы, такие, как SVM. Важным свойством метода является возможность получения компактных словарей, описывающих группы текстовых документов.

В будущем авторы предполагают использовать инкрементный алгоритм совместной кластеризации [33], который позволит, во-первых, ускорить работу предложенного метода, во-вторых, осуществлять классификацию в случае потока тестовых документов: тогда нужно будет заранее обучить классификатор для кластеров, включающих только известные размеченные данные, а тестовые данные в режиме “онлайн” будут размещаться по кластерам и классифицироваться. Кроме этого, интересно исследовать вариант предложенного метода с перекрывающимися кластерами [14].

СПИСОК ЛИТЕРАТУРЫ

1. *Pang B., Lee L.* Opinion Mining and Sentiment Analysis // Foundations and Trends® in Information Retrieval. 2008. V. 2. P. 1–135.
2. *Liu B.* Sentiment Analysis and Opinion Mining. San Rafael, California, USA: Morgan & Claypool Publishers, 2012.
3. *Chetviorkin I., Braslavskiy P., Loukachevitch N.* Sentiment Analysis Track at ROMIP 2011 // Computational Linguistics and Intellectual Technologies. 2012. V. 11(18). P. 739–746.
4. *Котельников Е.В., Клековкина М.В.* Определение весов оценочных слов на основе генетического алгоритма в задаче анализа тональности текстов // Программные продукты и системы. 2013. № 4. С. 296–300.
5. *Dhillon I. S.* Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning // Proc. 7th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining (KDD-2001). San Francisco, CA, USA, 2001. P. 269–274.
6. *Blinov P., Klekovkina M., Kotelnikov E., Pestov O.* Research of Lexical Approach and Machine Learning Methods for Sentiment Analysis // Computational Linguistics and Intellectual Technologies. 2013. V. 12(19). P. 51–61.
7. *Sebastiani F.* Machine Learning in Automated Text Categorization // ACM Computing Surveys. 2002. V. 34(1). P. 1–47.
8. *Kyriakopoulou A., Kalamboukis T.* Using Clustering to Enhance Text Classification // Proc. 30th Annual Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval. Amsterdam, Holland, 2007. P. 805–806.
9. *Williams G.K., Anand S.S.* Predicting the Polarity Strength of Adjectives Using WordNet // Proc. 3rd Intern. ICWSM Conf. San Jose, California, USA, 2009. P. 346–349.
10. *Kamps J., Marx M.J., Mokken R.J., de Rijke M.* Using WordNet to Measure Semantic Orientations of Adjectives // Proc. 4th Intern. Conf. on Language Resources and Evaluation. Lisbon, Portugal, 2004. P. 1115–1118.
11. *Chetviorkin I., Loukachevitch N.* Extraction of Russian Sentiment Lexicon for Product Meta-Domain // Proc. COLING 2012: Technical Papers. Mumbai, India, 2012. P. 593–610.
12. *Turney P.* Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews // Proc. Annual Meeting of the ACL. Stroudsburg, PA, USA, 2002. P. 417–424.
13. *Hartigan J.A.* Direct Clustering of a Data Matrix // J. American Statistical Association. 1972. V. 67(337). P. 123–129.
14. *Zha H., He X., Ding C., Simon H., Gu M.* Bipartite Graph Partitioning and Data Clustering // Proc. 10th Intern. Conf. on Information and Knowledge Management. Atlanta, Georgia, USA, 2001. P. 25–32.

15. *Голуб Дж., Ван Лоун Ч.* Матричные вычисления. М.: Мир, 1999. 548 с.
16. *Fu X., Guo Y., Guo W., Wang Z.* Aspect and Sentiment Extraction Based on Information-Theoretic Co-clustering // Proc. 9th Intern. Symp. on Neural Networks. Pt II. Lecture Notes in Computer Science. Shenyang, China, 2012. V. 7368. P. 326–335.
17. *Raison K., Tomuro N., Lytinen S., Zagal J.P.* Extraction of User Opinions by Adjective-Context Co-clustering for Game Review Texts // Proc. 8th Intern. Conf. on NLP. Kanazawa, Japan, 2012. P. 289–299.
18. *Li T., Zhang Y., Sindhwani V.* A Non-negative Matrix Tri-factorization Approach to Sentiment Classification with Lexical Prior Knowledge // Proc. 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. Singapore, 2009. P. 244–252.
19. *Sindhwani V., Melville P.* Document-Word Co-Regularization for Semi-supervised Sentiment Analysis // Proc. 8th IEEE Intern. Conf. on Data Mining (ICDM '08). Pisa, Italy, 2008. P. 1025–1030.
20. *Frigui H., Nasraoui O.* Simultaneous Clustering and Dynamic Keyword Weighting for Text Documents // Survey of Text Mining / Ed. M. Berry. N.Y.: Springer, 2004. P. 45–70.
21. *Holland J.* Adaptation in Natural and Artificial Systems. Ann Arbor, MI, USA: University of Michigan Press, 1975.
22. *Рутковская Д., Пилиньский М., Рутковский Л.* Нейронные сети, генетический алгоритмы и нечеткие системы. М.: Горячая линия-Телеком, 2007. 452 с.
23. *Li J., Zheng R., Chen H.* From Fingerprint to Writeprint // Communications ACM. 2006. V. 49(4). P. 76–82.
24. *Oliveira L.S., Sabourin R., Bortolozzi F., Suen C.Y.* Feature Selection Using Multiobjective Genetic Algorithms for Handwritten Digit Recognition // Proc. 16th Intern. Conf. on Pattern Recognition. Quebec, Canada, 2002. P. 568–571.
25. *Vafaie H., Imam I.F.* Feature Selection Methods: Genetic Algorithms vs. Greedy-like Search // Proc. 3rd Intern. Fuzzy Systems and Intelligent Control Conf. Louisville, KY, USA, 1994.
26. *Abbasi A., Chen H., Salem A.* Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums // ACM Transactions on Information Systems. 2008. V. 26(3). P. 1–34.
27. *Das A., Bandyopadhyay S.* Subjectivity Detection using Genetic Algorithm // 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. Lisbon, Portugal, 2010.
28. *Paltoglou G., Thelwall M.* A Study of Information Retrieval Weighting Schemes for Sentiment Analysis // Proc. 48th Annual Meeting Association for Computational Linguistics. Uppsala, Sweden, 2010. P. 1386–1395.
29. *Зализняк А. А.* Грамматический словарь русского языка. Изд. 5-е, испр. М.: Аст-пресс, 2008.
30. *Kohavi R.* A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection // Proc. 14th Intern. Joint Conf. on Artificial Intelligence. Quebec, Canada, 1995. V. 2(12). P. 1137–1143.
31. *Chang C.-C., Lin C.-J.* LIBSVM: a Library for Support Vector Machines // ACM Transactions on Intelligent Systems and Technology. 2011. V. 2(3). P. 1–27.
32. *Pang B., Lee L., Vaithyanathan S.* Thumbs up? Sentiment Classification Using Machine Learning Techniques // Proc. Conf. on Empirical Methods in Natural Language Processing. Philadelphia, Pennsylvania, USA, 2002. P. 79–86.
33. *Pensa R. G., Ienco D., Meo R.* Hierarchical Co-clustering: Off-line and Incremental Approaches // Data Mining and Knowledge Discovery. 2014. V. 28(1). P. 31–64.